JUMS

Volume 11, Issue 1, April 2026

JUNIOR MANAGEMENT SCIENCE

Editor:
DOMINIK VAN AAKEN

Advisory Editorial Board:
FREDERIK AHLEMANN
JAN-PHILIPP AHRENS
THOMAS BAHLINGER
MARKUS BECKMANN
SULEIKA BORT
ROLF BRÜHL
KATRIN BURMEISTER-LAMP
CATHERINE CLEOPHAS
NILS CRASSELT
BENEDIKT DOWNAR
KERSTIN FEHRE
MATTHIAS FINK
DAVID FLORYSIAK
GUNTHER FRIEDL
MARTIN FRIESL
FRANZ FUERST
WOLFGANG GÜTTEL
NINA KATRIN HANSEN
ANNE KATARINA HEIDER
CHRISTIAN HOFMANN
SVEN HÖRNER
STEPHAN KAISER
NADINE KAMMERLANDER
ALFRED KIESER
ALEKSANDRA KLEIN
NATALIA KLIEWER
DODO ZU KNYPHAUSEN-AUFSESS
SABINE T. KÖSZEGI
ARJAN KOZICA
CHRISTIAN KOZIOL
MARTIN KREEB
WERNER KUNZ
HANS-ULRICH KÜPPER
MICHAEL MEYER
JÜRGEN MÜHLBACHER
GORDON MÜLLER-SEITZ
J. PETER MURMANN
ANDREAS OSTERMAIER
BURKHARD PEDELL
ARTHUR POSCH
MARCEL PROKOPCZUK
TANJA RABL
SASCHA RAITHEL
NICOLE RATZINGER-SAKEL
ASTRID REICHEL
KATJA ROST
THOMAS RUSSACK
FLORIAN SAHLING
MARKO SARSTEDT
ANDREAS G. SCHERER
STEFAN SCHMID
UTE SCHMIEL
CHRISTIAN SCHMITZ
MARTIN SCHNEIDER
MARKUS SCHOLZ
LARS SCHWEIZER
DAVID SEIDL
THORSTEN SELLHORN
STEFAN SEURING
VIOLETTA SPLITTER
ANDREAS SUCHANEK
TILL TALAULICAR
ANN TANK
ANJA TUSCHKE
MATTHIAS UHL
CHRISTINE VALLASTER
PATRICK VELTE
CHRISTIAN VÖGTLIN
BARBARA E. WEISSENBERGER
ISABELL M. WELPE
HANNES WINNER
THOMAS WRONA
THOMAS ZWICK

Christopher Dudel, Runtime-Sensitive Learned Operator Selection in ALNS: Testing Improvements to Adaptive Operator Selection while Optimizing 1

David Schopp, Die Rolle von Corporate Digital Responsibility im digitalen Controlling 27

Leonie Terhardt, Textauswertung in der Finanzanalyse: Eine Ereignisstudie zum Sentiment in Earnings Calls 43

Moritz Praetz, The Impact of Biodiversity Risk on Banks' Credit Default Swap Spread Changes 74

Sebastian Schmidt, Kausale Inferenz Unter Anwendung Von Double Machine Learning: Oregon Health Insurance Experiment 107

Susanne Rautzenberg, The Dark Side of Employer Branding – Aesthetic Labour and Employer Attractiveness in the Beauty and Cosmetics Industry 139

Justus Scharfstädt, Understanding Participation Trends in the Avani 5G Patent Pool: A Descriptive Analysis 164

Philipp Iversen, Generative AI-Enabled Music Generation in Marketing and Consumer Response 181

Ege Özkul, Predicting Stock Market Trends Using Convolutional Neural Networks: A Deep Learning Approach 195

Felix Achim Bautz, Relative Performance-Messung als Instrument der Unternehmenssteuerung 227

Published by Junior Management Science e.V.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

ISSN: 2942-1861

Junior Management Science 11(1) (2026) 1-26

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Runtime-Sensitive Learned Operator Selection in ALNS: Testing Improvements to Adaptive Operator Selection while Optimizing Runtime

Christopher Dudel

Technical University of Munich

Abstract

We propose and test two variations of the Adaptive Large Neighborhood Search (ALNS) meta-heuristic: First, we add time sensitivity to the operator selection scheme to optimize the ALNS for both solution quality and runtime. We reward comparatively slow operators with reduced rewards for finding improvements. This ensures that the meta-heuristic is slowed down less by operators which consistently find good solutions but take long to do so. Secondly, we replace the Adaptive Layer with a Learned Operator Selection Policy trained via Deep-Q Learning. The training takes both solution quality and operator runtime into account. We test our algorithms against classic ALNS as well as random operator selection. We perform an analysis of how operator portfolios affect performance. Our chosen problem domain is the Capacitated Vehicle Routing Problem with 100 to 400 customer nodes.

Keywords: adaptive large neighborhood search; vehicle routing; optimization; logistics; deep learning

1 Introduction

The Adaptive Large Neighborhood Search (ALNS) is a meta-heuristic proposed by Ropke and Pisinger (2006) which is commonly used to solve Operations Research (OR) problems such as the Capacitated Vehicle Routing Problem (CVRP). Recently, the effectiveness of the main feature of ALNS, the adaptive operator selection, has been called into question (Turkeš et al., 2021). This study identifies potential improvements to ALNS.

ALNS calculates an operator's selection probability based on the value function of explored solutions. Finding, for example, the lowest cost for a routing problem is the end goal of an optimization heuristic. But if finding the optimal solution value were the only goal, one could use exact solvers for any Combinatorial Optimization (CO) problem. The motivation to develop and use heuristic algorithms is to reduce runtimes to manageable levels while still finding acceptable solutions. Following this logic, ALNS can benefit from being able to choose its sub-heuristics based on both their solution qualities and the runtimes it takes to explore them. We sample the available literature already implementing runtime-sensitive ALNS and analyze the impact of operator portfolios on the effectiveness of the runtime-sensitive Adaptive Large Neighborhood Search (tALNS).

Operator runtime is only one additional observation that can be used to better inform the weight update procedure. Yet, building a weight update function that correctly balances the inputs becomes more complicated with each new information. Artificial Neural Networks (ANNs) can be used as function approximators of complex in-out-relationships (Ferrari & Stengel, 2005). (Boualamia et al., 2023) successfully replace the weight update formula with an ANN informed only by the operator reward history within the optimization run.

ALNS goes through a warm-up period, in which the avail-

I would like to thank Dr. Christina Liepold for her exceptional tutor- and mentorship. Her constant support and great advice were immeasurably helpful to this work. I sincerely thank Prof. Dr. Maximilian Schiffer for getting me interested in optimization algorithms and the opportunity to write my thesis, as well as the pragmatic and exciting approach to thesis-writing he promotes. Lastly, I am deeply grateful to Prof. Dr. Josef D. for supporting my studies.

DOI: https://doi.org/10.5282/jums/v11i1pp1-26

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

able operators are sampled, and operator selection probabilities start to converge towards the actual operator performance (Bongiovanni et al., 2022). Due to the stochastic nature of the operator heuristics, this warm-up can last for a significant portion of the meta-heuristic run. In small problem instances, the run might be over before adequate probabilities have formed. In large problems, operator weights decay towards the lower minimum as finding improvements becomes scarce. These are inherent downsides of the operator selection only being informed by information from previous iteration outcomes in one optimization run (Reijnen et al., 2022). By informing the selection with data from many runs and by training an operator selection policy represented by an ANN with Deep-Q Learning (DQL), we can provide a better base for the operator selection. As this learned policy replaces the adaptive layer, we call this algorithm the Learned Large Neighborhood Search (LLNS).

We differ from previous implementations of tALNS and LLNS by exploring large CVRP problem instances with up to 400 customer nodes. To the best of our knowledge, we provide the most extensive comparison of ALNS and tALNS up to date. We analyze how changing operator portfolios and the inclusion of a Time-Sensitivity Factor affect the performance of the two algorithms. In the field of LLNS, this study is - to the best of our knowledge - the first to optimize operator selection for both solution quality and runtime.

2 State of the Art

In the following, we discuss the literature related to our research: We seek to solve the CVRP - a classic vehicle routing problem domain - utilizing the ALNS meta-heuristic. We pay special attention to approaches that aim to optimize both runtime and solution quality of algorithms. We further discuss how methods from the field of Machine Learning (ML) are used to improve proven OR techniques.

2.1 Capacitated Vehicle Routing Problem

The Capacitated Vehicle Routing Problem (CVRP) is a logistics optimization problem concerned with finding the optimal assignment of a given set of customer nodes to a fleet of vehicles. The vehicles are limited in the number of customers that can be served in one route, for example, via a given carrying capacity. See Godinho et al. (2008) for an in-depth definition and a mathematical formulation of the CVRP. Routing problems have been explored in OR since the 1950s (Dantzig & Ramser, 1959). Vehicle Routing Problems (VRPs) are often expanded with additional details and constraints to make them more applicable to real-world settings, such as heterogeneous fleets (Kucukoglu et al., 2021). Nevertheless, even the basic CVRP is NP-hard and time-consuming to solve via exact methods (Hochba, 1997). Therefore, heuristic approaches are often employed to find acceptable solutions to CVRPs (Gendreau & Potvin, 2010). Heuristic approaches often suffer from getting trapped in local optima. *Meta-heuristics* are a group of algorithms designed to find global optima by combining local improvement heuristics with a higher-level search strategy. They are a powerful method to solve complex CO problems like the CVRP (Gendreau & Potvin, 2010, p. ix).

2.2 Adaptive Large Neighborhood Search

One popular meta-heuristic often applied to routing problems is the Adaptive Large Neighborhood Search (ALNS). (Ropke & Pisinger, 2006) propose ALNS as an extension to the Large Neighborhood Search (LNS) meta-heuristic developed by Shaw (1998). LNS iteratively uses a local search heuristic to destroy and repair an existing solution to find improvements to the objective value. ALNS expands on this approach by allowing sets of destroy and repair heuristics instead of reusing one heuristic in each iteration. A set of operator weights governs which operators are selected for a given iteration. These weights **adapt** in response to the historical performance of the operators within the optimization run. This increases the algorithm's robustness by enabling self-calibration towards using the most efficient operators for a given problem (Pisinger & Ropke, 2007). Section 3.1 and Algorithm 1 give a detailed explanation of the ALNS.

ALNS is successfully used to solve large-scale routing and scheduling problems (Gendreau & Potvin, 2010, pp. 409f.). For an in-depth survey of ALNS applications published until mid-2021, see Windras Mara et al. (2022). They list 228 implementations and find a growing trend for the yearly number of published ALNS studies. This shows that ALNS continues to be seen as an effective optimization framework. (Turkeš et al., 2021) perform a meta-study on ALNS focused on quantifying the average solution quality improvement gained by implementing an adaptive operator selection process. They ask authors of ALNS implementations to re-solve their optimization problems with the same operator portfolios but without weight updates, leading to a random operator selection. They call this version of the meta-heuristic non-adaptive Large Neighborhood Search ((¬)ALNS). (Turkeš et al., 2021) find that, on average, the solution quality of ALNS only improves by 0.14% compared to (¬)ALNS. This relatively low improvement results in the question under which circumstances implementing ALNS over (¬)ALNS is worth the added complexity and whether an improved operator selection mechanism exists (Johnn et al., 2023).

(Turkeš et al., 2021) note two positive outliers in their comparison between ALNS and (¬)ALNS: (Thomas & Schaus, 2018) implement ALNS as a general purpose black box optimizer, set up to be able to solve a wide range of OR problems. For this purpose, they use 30 destroy and 36 repair-heuristics, many more than are commonly pre-selected (Windras Mara et al., 2022). Furthermore, (Thomas & Schaus, 2018) adjust operator weights based on solution quality improvements and operator runtimes. Compared to random operator selection, their ALNS improves solution quality by 15.46%. (Kiefer et al., 2017) manage to improve solution quality by 6.54% over (¬)ALNS. They also utilize more-than-average operators with ten destroy- and eight repair heuristics. Additionally, they also use runtime-sensitivity in their

operator weight update scheme. Both these studies are discussed in more detail in sections 2.2.2 and 2.2.3.

2.2.1 Motivating Runtime-Sensitive Meta-Heuristics

The primary objective of optimization algorithms is to reach a high-quality solution for a given problem. Secondary objectives exist, such as solving within a reasonable time frame, robustness against varying problem characteristics, or low outcome variability (Bengio et al., 2021). Especially in commercial operations, obtaining a good solution fast is often preferred to finding a slightly better solution in a much longer time frame (Ahuja et al., 2002; Wu et al., 2017). Ropke and Pisinger state that their ALNS finds improvements to more than 50% of previous best-known solutions and that it does so "in a reasonable amount of time" (Ropke & Pisinger, 2006, p. 470). The adaptive layer's objective is to choose heuristics that reliably find improvements to the current solution. However, heuristics differ not only in their solution quality but also in their runtime (Ahuja et al., 2002).

(Ahuja et al., 2002) discuss the importance of runtime optimization in large-scale neighborhood searches. Yet few studies are implementing ALNS with an explicit focus on fast solving times. The literature research by Windras Mara et al. (2022) notes the importance of selecting operators that can explore large neighborhoods time-efficiently but does not identify literature that implements time-sensitivity. (Turkeš et al., 2021) focus their meta-analysis of the adaptive layer only on its impact on the solution quality, but note that some authors (analyzed in Table 1) have modified the weight adjustment formula to incorporate operator runtimes (Adulyasak et al., 2014; Canca et al., 2018; Gullhav et al., 2017; Thomas & Schaus, 2018; Wang et al., 2012; Wu et al., 2017). The following section presents a literature review of runtime-sensitive ALNS, focusing on how operator runtime is taken into account in the adaptive layer and on how algorithms can be compared in multi-objective settings where both solution quality and runtime are optimized.

2.2.2 Runtime-Sensitive Operator Selection

Sophisticated heuristics like, for example, *regret_repair* (see Section 3.1.1) - which has to keep track of and repeatedly reevaluate multiple insertion points for each non-inserted node - exceed the runtime of quicker heuristics like *greedy_repair* by a considerable margin. Consider two theoretical heuristics: *fast-repair* and *sophisticated-repair*. *Fast-repair* finds an improvement every tenth iteration. *Sophisticated-repair* finds an improvement in every iteration but takes 100 times longer to complete than *fast-repair*. ALNS would prioritize *sophisticated-repair*, as its higher rewards per iteration would quickly make its operator weight outgrow the weight of *fast-repair*. Given a fixed number of iterations, as is the case in Ropke and Pisinger (2006), this would be the correct approach to reach the best solution quality. However, given a fixed solving time, selecting *fast-repair* more often is preferable.

Table 1 lists the existing research into enhancing ALNS with runtime-sensitive rewards. Note that the lower three papers do not use the ALNS-framework, but the similar Self-Adapting Large Neighborhood Search (SA-LNS). Instead of utilizing destroy- and repair-heuristics, this meta-heuristic iteratively improves a solution via relaxing and re-tightening constraints, mainly in the domain of machine scheduling problems (Laborie & Godard, 2007).

Pairwise or Individual Operator Selection

In classic ALNS, destroy- and repair-operator weights are updated independently of each other (Ropke & Pisinger, 2006). This has the downside of not being able to tell which operator was responsible for the success or failure of the rewarded iteration and whether there are certain destroy- and repair-pairings that work particularly well. This downside is sometimes mitigated by rewarding and selecting destroy-repair-operator-pairs. However, this approach leads to having fewer data points to evaluate since each pair is chosen less often than each independent operator would be (Turkeš et al., 2021).

The underlying problem is further exacerbated in runtime-sensitive ALNS: the chosen destroy-operator can significantly influence the runtime of the repair-operator, especially if the degree of destruction differs between destroy-operators. Within the sampled literature, only Zhang and Yang (2021) evaluate operators in pairs. This is consistent with the works analyzed by Turkeš et al. (2021), out of which 81% use independent operator evaluation. The algorithms proposed in this paper follow this majority. As we utilize an extensive operator portfolio - leading to an excessive number of operator combinations - we select destroy and repair operators individually.

Table 1: Runtime-Sensitive Operator Selection: State of the Art

Reference	Weight-Update Formula for Updated Operator	Explanation
ALNS		
Adulyasak et al. 2014	$(1-\alpha) \times w_{b-1} + \alpha \times \frac{\sum_{i=1}^{\beta} \phi_i}{\Theta \times tf}$ $tf = \frac{\varnothing_b t_{fastest_operator}}{\varnothing_b t}$	After each batch b of iterations, operator weights w are updated via the gained rewards ϕ. The time factor tf is calculated by dividing the average runtime t within the last batch of the fastest operator with the average runtime of the operator which is currently updated.
Canca et al. 2018	$(1-\alpha) \times w_{b-1} + \alpha \times \frac{\sum_{i=1}^{\beta} \phi_i}{\Theta \times tf}$ tf is unspecified	An unexplained factor tf representing the computational effort required by the operator is used to scale the gained rewards after each batch.
Gullhav et al. 2017	$(1-\alpha) \times w_{b-1} + \alpha \times \frac{\sum_{i=1}^{\beta} rt_i}{\Theta}$ $rt_i =$ $max(\phi_{rej.},\ \phi_i \times min(1, \frac{T_{avg}}{t_i \times \omega}))$	Each reward is scaled by a normalized factor derived by dividing the average operator CPU time by the observed operator CPU time. The observed CPU time is further scaled by a meta-parameter ω (tuned in range 0.5-1.5). The reward can only be lowered for slower-than-average operators. Rewards can also never be lower than the reward gained for finding a rejected solution $\phi_{rej.}$.
Kiefer et al. 2017	$(1-\alpha) \times w_{b-1} + \alpha \times \frac{\sum_{i=1}^{\beta} \phi_i}{\Theta \times tf}$ $tf = \frac{\varnothing t_{reference_operator}}{\varnothing_b t}$	Runtime-factor tf is calculated by dividing the average runtime of a pre-chosen reference operator by the average runtime of the updated operator within the last batch. tf is only present for repair-operators.
Wu et al. 2017	$(1-\alpha) \times w_{i-1} + \alpha \times \frac{\phi_i}{tf}$ tf is unspecified	The reward gained for finding a best or accepted solution is scaled with an undisclosed time cost tf.
Zhang and Yang 2021	$(1-\alpha) \times w_{i-1} + \alpha \times \frac{\phi_i}{tf_i}$ tf_i is runtime of iteration i	The adaptive layer rewards combinations of destroy- and repair-operators. Rewards are scaled with the execution time of the iteration.
SA-LNS		
Laborie and Godard 2007	$(1-\alpha) \times w_{i-1} + \alpha \times \phi_i$ $\phi_i = \frac{\triangle c_i}{\triangle t_i}$	Probabilities for relaxation and completion strategies are updated each iteration via a ratio derived from the solution improvement (if any) $\triangle c_i$ and the CPU cycle time $\triangle t_i$. How the CPU time relates to the solution changes or whether either of the two inputs is normalized is not explained.
Thomas and Schaus 2018	$(1-\alpha) \times \frac{L}{T} \times T_o + \alpha \times L_o$ $L_o = \frac{\sum_{i^*}^{i} \triangle c_{io}}{\sum_{i^*}^{i} \triangle t_{io}}$ $T_o = \frac{\sum_{0}^{i} \triangle c_{io}}{\sum_{0}^{i} \triangle t_{io}}$ L and T likewise but over all operators	An expansion on SA-LNS updating weights each iteration, but over a rolling time window starting slightly before the iteration of the current best-known solution (i^*). Within this window, the local efficiency of each operator L_o is calculated. The total operator efficiency T_o, which is taken over all iterations, serves to smooth out the weight changes. The total efficiency is further scaled by the ratio between local and total efficiency over all operators (L and T).
Wang et al. 2012	$(1-\alpha) \times w_{i-1} + \alpha \times \frac{\phi_i}{tf_i}$ tf_i is the computational time in the current iteration	No information is given which operations within the iteration count towards tf.

α: weight update factor. β: size of batch. Θ: operator appearances in batch.

Table 2: Algorithm Acceptance Criteria: State of the Art

Reference	Stopping Criterion	Comparison
ALNS		
Adulyasak et al. 2014	100 iter. no new accepted / 300 iter. no new best*	Both solution quality and runtimes are reported independently. Additionally, average operator performances and runtimes are reported.
Canca et al. 2018	3,600 seconds	Algorithms are compared by best solution cost reached within stopping criterion.
Gullhav et al. 2017	maximum iterations	The maximum iterations are set so that an average run takes slightly longer than 15 minutes. Average solution GTOs are compared at 5, 10 and 15 minutes.
Kiefer et al. 2017	480 seconds	Report the average cost over multiple runs, and best cost found for each problem instance. They judge that their proposed algorithm outperforms a competitor because it finds better solutions than the competitor in more instances.
Wu et al. 2017	maximum iterations	Stopping criterion is set dependent on problem size, with larger problems having less iterations to limit solving times. Both time and solution quality are reported and compared qualitatively. The authors justify achieving worse solutions than comparable algorithms because they limited themselves to runtimes of one hour.
Zhang and Yang 2021	$I_{max} = 25,000 + 50\times$ instance size	Solution quality and runtime are independently reported, but not compared.
SA-LNS		
Laborie and Godard 2007	max. consecutive non-improving iter.	Stopping criterion is dependent on problem size. Algorithms are judged solely by final solution quality. Solving times are not reported.
Thomas and Schaus 2018	240 seconds	Present a figure showing average solution costs at solving times for their proposed algorithm and a competitor. Their algorithm outperforms the competitor at all times during the 240 seconds.
Wang et al. 2012	max. consecutive non-improving iter. / 600 seconds*	Report the solution gap to optimum (calculated exactly with CPLEX). If the gap to optimum is zero, a time savings in seconds to the runtime of CPLEX is reported.

I_{max} refers to the maximum number of iterations.

*Whichever is reached first

2.2.3 Algorithm Comparison Criterion

The challenge of multi-objective comparisons between heuristics appears again when comparing the overall performance of meta-heuristics. The fundamental question that needs to be answered is: how much runtime is a given increase in solution quality worth?

Table 2 shows how the authors exploring runtime-sensitive ALNS and SA-LNS compare different meta-heuristic implementations. Comparing solution quality after a fixed runtime is the preferred method in the literature sampled. However, some drawbacks are associated with this approach, which are discussed in Section 4.2. Instead, we opt to inverse the two optimization objectives and compare the elapsed runtime upon reaching a fixed solution quality.

2.3 Combining Machine Learning with Meta-Heuristics

Recently, multiple approaches to utilize Machine Learning (ML) for solving classical CO problems have been proposed (Bengio et al., 2021; Karimi-Mamaghan et al., 2022; Reijnen et al., 2022). The following section starts with an overview of the field of ML in CO and proceeds with a deep-dive into the literature studying a learned operator selection policy for ALNS.

2.3.1 Machine Learning in Combinatorial Optimization

Combinatorial Optimization (CO) problems are often time-intensive to solve, and the field is constantly searching for improved algorithms to optimize hard problems in less time. One improvement step was the development of powerful meta-heuristics. A current research direction is to exploit ML-techniques, which have recently significantly improved in performance and spread of adoption. (Bengio et al., 2021) present an overview into the combination of CO and ML.

A straightforward approach to utilize ML for CO is to directly learn solutions with a given OR problem as input: (Nazari et al., 2018) do so for the VRP, finding solutions for instances with up to 100 nodes. (Kool et al., 2018) train an Attention Model to optimize the Traveling Salesman Problem (TSP) and CVRP. Currently, these direct approaches suffer from exceedingly long runtimes when scaling to larger instances (Reijnen et al., 2022).

A second idea is to combine a proven optimization approach with a specialized learned component. Several authors (Song et al., 2020; Sonnerat et al., 2021; Wu et al., 2021; Zhou et al., 2023) enhance exact solvers of Linear and Mixed Integer Problems with ML, for example by selecting constraints to relax via a learned policy. (Da Costa et al., 2021) take a heuristic approach and pick nodes for 2-opt moves via Deep Reinforcement Learning (DRL).

Meta-heuristics have proven efficient at optimizing large-scale OR problems (Gendreau & Potvin, 2010). Increasingly, ML techniques have been combined with meta-heuristics to create powerful optimization algorithms (Karimi-Mamaghan et al., 2022; Oliva et al., 2021). A review of the current state of the art for meta-heuristics enhanced by ML can be found in Karimi-Mamaghan et al. (2022). For example, (Joe & Lau, 2020) combine DRL with Simulated Annealing (SA) for the Dynamic CVRP. The LNS by Shaw (1998) has been a popular meta-heuristic to utilize around a ML operator: (Gao et al., 2020) chose promising VRP neighborhoods by using a trained agent as the destroy operator. (Wu et al., 2022) learn an LNS improvement heuristic for TSP and CVRP. (Hottung & Tierney, 2022) propose a Neural Large Neighborhood Search with a learned improvement heuristic for CVRP, Split Delivery Vehicle Routing Problem, and Capacitated Team Orienteering Problem.

Partially motivated by the findings of Turkeš et al. (2021), another approach has emerged which seeks to improve the adaptive layer of ALNS with a learned operator selection. The meta-heuristic is modeled as a Marcov Decision Process (MDP), where each step entails selecting the most efficient operators for the current iteration (Johnn et al., 2023). We identify four studies (Bongiovanni et al., 2022; Boualamia et al., 2023; Johnn et al., 2023; Reijnen et al., 2022) adapting this approach and explore them in Section 2.3.2. The field of selecting the best algorithm to solve a given combinatorial optimization problem is called Automated Algorithm Selection (Kerschke et al., 2019). Reviews can be found in Kerschke et al. (2019) and Karimi-Mamaghan et al. (2022). We further identify two papers from this field that do not utilize the ALNS-framework but are close enough in their approach to warrant mentioning: (Chen & Tian, 2018) propose an actor-critic neural network to rewrite feasible solutions of Job Scheduling and VRP problems with up to 100 customers. Two policies are trained: A region-picking policy that chooses which features of the previous solution to remove and a rule-picking policy that determines how the next solution is created. (Lu et al., 2020) propose a novel heuristic framework that iteratively improves a feasible solution. Operators are split into two pools: perturbation and improvement. A meta-controller decides which pool to choose from for the next iteration. An attention network is used to train the selection of the improvement operators, while perturbation operators are chosen randomly. They use this framework to find solutions for CVRP-instances of size 100.

2.3.2 Learned Large Neighborhood Search

The approach by Boualamia et al. (2023) stays comparatively close to classic ALNS: Instead of training a policy over multiple optimization runs, they opt to replace the ALNS weight update and roulette wheel selection scheme with an epsilon-greedy policy learned anew within each optimization run. The Q-Table, which effectively replaces the operator weights, is updated via rewards based on the acceptance check outcome of each iteration, as is the case in ALNS. They achieve exploration of destroy-repair-pairs by starting with a chance of $\varepsilon = 0.6$ to select a random pairing. This chance exponentially decays to exploit pairings that have proven efficient. As their approach requires no training, it can be applied to broad ranges of problem instances, demonstrated by optimizing CVRP instances between 45 and 1200 nodes.

(Bongiovanni et al., 2022) optimize the Autonomous

Ridesharing Problem. They build a dataset of 1.5 million transitions, which they use for offline training of an ensemble of Random Forrest classification predictors. Each pair of destroy- and repair operators is associated with a Random Forrest model, which predicts whether the pair will likely find an improvement to the active solution. A roulette wheel is then used to draw from all pairs which are deemed suitable. The authors utilize a large feature portfolio, including detailed domain-specific observations such as the current average state of charge of all vehicles. The predicted reward is the improvement to the solution quality after using the operators for one iteration.

(Johnn et al., 2023) exploit extensive knowledge about the search state to better inform the operator selection. They train a selection policy via Deep-Q Learning, testing both a Multi-Layer Perceptron and a Graph Neural Network as function approximators. To inform the selection, they translate both the CVRP instance and the current solution state - represented by the existing routes - into graphs, which they pass as features. A single reward is given at the end of each run, calculated as the improvement in solution quality from the initial to the best solution. Training is done on 20-node CVRP instances, with results showing that the trained policies still outperform ALNS in larger problems. The authors find that the learned policy leads to greater improvements over ALNS with growing operator portfolios.

(Reijnen et al., 2022) take a different approach to Johnn et al. (2023) and Bongiovanni et al. (2022) in that they only consider domain-agnostic features, such as the number of iterations since the current best solution was found. They use Deep Reinforcement Learning to train a pairwise operator selection policy. After 32 hours of training on instances of size 20, their agent is able to outperform ALNS on instances of size 100. Their chosen domain for the comparison is the Time-Dependent Orienteering Problem with Time Windows.

Table 3 gives an overview of the relevant features of the sampled studies.

Table 3: Learned Large Neighborhood Search: State of the Art

Reference	Learning Technique	Num. Ω^-/Ω^+	Pairwise	Instance Size	Dimension	Num. Specific / Agnostic Features	Rewards
Bongiovanni et al., 2022	Random Forrest Ensemble	3 / 3	✓	Autonomous Ridesharing	24.000 requests over 7 days	40 / 6	Objective Improvement in Iteration
Boualamia et al., 2023	Q-Learning with Epsilon-Greedy Policy	4 / 2	✓	CVRP	45 - 1200	- / -	ALNS rewards
Johnn et al., 2023	DQL, Multi-Layer Perceptron & Graph Neural Network	12 / 2	X	CVRP	20 - 100	Instance graph & route graphs / -	Solution quality increase from first to last iteration
Reijnen et al., 2022	Deep Reinforcement Learning	4 / 3	✓	Orienteering Problem	20 - 100	- / 8	ALNS rewards
This LLNS	DQL, C51	5* / 7	X	CVRP	100 - 399	- / 10	Multi-Objective: solution quality & runtime

Ω^- / Ω^+ refers to the sets of destroy / repair operators. *Only 1 destroy operator is used in Algorithm Comparison.
Pairwise refers to whether destroy and repair operators are chosen as a pair, or independently.
Specific / Agnostic refers to whether or not a feature is independent from the problem domain.
C51 is a variation of Q-Learning where the expected Q-value is represented by a distribution (Bellemare et al., 2017).

2.4 Contribution

The presented literature review shows that significant efforts have been made to improve the ALNS adaptive layer. Nevertheless, there exists a need for further study: in the area of runtime-sensitive ALNS, we note that none of the authors in our review put the focus of their work onto the time-sensitivity. Section 4.6 will present an in-depth comparison of ALNS and tALNS, including discussing an edge-case where tALNS can perform worse. We prove that a tuned Time-Sensitivity Factor ω, previously only used by Gullhav et al. (2017), is vital to the performance of tALNS. No analysis of this kind exists in the sampled literature.

In the case of LLNS, we differentiate on three levels: First, we expand the instance size range from the commonly used VRP-100 up to 400 nodes, with 700+ node instances used to test generalization. Furthermore, we focus on using domain-agnostic observations. This opens up the possibility to use the LLNS-framework outside of CVRP. Lastly, to the best of our knowledge, this study is the first to combine runtime-sensitivity and learned operator selection. This presents a significant challenge for training, as training results depend on a complex multi-objective reward function, which has to be well-tuned to allow learning a time-efficient operator selection.

3 Methodology

As we evaluate two potential improvements to ALNS, the Methodology section is split into two parts. Section 3.1 explains the process of classic ALNS and our proposal of tALNS. Section 3.2 details how we learn the operator selection policy with ML

3.1 Runtime-Sensitive Adaptive Large Neighborhood Search

To make the results of this study comparable to other uses of ALNS, an effort was made to not deviate from a standard implementation of ALNS wherever feasible. We use the most prominent Adaptive Mechanism, Acceptance Criterion, and Termination Criterion as surveyed by Windras Mara et al. (2022). Their literature review shows that, out of 251 surveyed ALNS implementations, 99% use a Roulette Wheel Adaptive Mechanism, 81% use a Metropolis Acceptance Criterion (such as SA) and 75% use a maximum number of iterations as the Termination Criterion. Figure 1 shows how the solution of a small CVRP-instance might evolve over the course of an ALNS-run.

Algorithm 1 gives an overview of the process of the metaheuristic. All relevant details are explained in depth within this chapter.

Central to Algorithm 1 is a solution S. It is initialized via a sweep operation, sorting nodes into feasible routes based on their angular position. Depending on the problem size and layout, sweep tends to find solutions between 25% and 10% distance to the optimal solution for our instances. Sampling the sweep results shows it performs better in problems that place the depot at the center. In addition to the sweep, we run one iteration each for every available repair heuristic. This is done to be able to calculate a starting temperature sat_0 for SA (Kirkpatrick et al., 1983).

This solution-initialization, which is marked as *sweep+* in Algorithm 1, leads to the creation of three tracked solutions: The active solution S, the best-known solution S^*, and an accepted solution S^a. They allow performing an acceptance check after each iteration i utilizing SA. The probability of accepting a worse solution is governed by the SA temperature sat. Section 3.1.2 explains the SA-setup.

ALNS differentiates from LNS by utilizing sets of destroy- and repair operators Ω^- and Ω^+. In each iteration, a destroy- and a repair operator $o^{-/+}$ are chosen via a roulette wheel selection dependent on operator weights w. After the repair phase, a local optimization is performed on the active solution. Section 3.1.1 explains the operators and the local optimization.

Algorithm 1: Algorithm for the (time-sensitive) Adaptive Large Neighborhood Search

Inputs: problem-instance $\mathscr{I}$, meta-parameters $\mathscr{P}$
Initialization:
S^*, sat $\longleftarrow$ sweep+_initialize($\mathscr{I}$)
w $\longleftarrow$ weight_initialize($\mathscr{P}, \Omega^-, \Omega^+$)
$S^a = S^*$
Iterations:
for $i = 1, \ldots, I_{max}$ **do**
  $S = S^a$
  o^-, o^+ $\longleftarrow$ choose(w)
  S, t^- $\longleftarrow$ destroy(S, o^-)
  S, t^+ $\longleftarrow$ repair(S, o^+)
  S $\longleftarrow$ local_optimization(S)
  S^*, S^a, ρ $\longleftarrow$ acceptance_check(S, S^*, S^a, sat)
  Φ, $ct^{-/+}, \Theta^{-/+}$ $\longleftarrow$ score_update(Φ, ρ, t^-, t^+)
  if i % β == *0* **then**
    w $\longleftarrow$ weight_update(w, Φ, ct, Θ)
    Φ, ct, $\Theta = 0$ for each o
return S^*

Based on the result of the acceptance check ρ, the active operators scores Φ are increased by a reward value of ϕ. Furthermore, we keep track of each operator's number of appearances Θ and cumulative runtime ct per batch. The operator weights are updated after a fixed batch b of iterations. The iterations within the next batch will select their operators based on the updated weights. Section 3.1.3 explains the weight update mechanism. Finally, after the stopping criterion is reached - in this case, a maximum number of iterations - the currently best-known solution is presented as the final result.

Based on the result of the acceptance check ρ, the active operators scores Φ are increased by a reward value of ϕ. Furthermore, we keep track of each operator's number of appearances Θ and cumulative runtime ct per batch. The operator weights are updated after a fixed batch b of iterations. The iterations within the next batch will select their operators based on the updated weights. Section 3.1.3 explains the weight update mechanism. Finally, after the stopping criterion is reached - in this case, a maximum number of iterations - the currently best-known solution is presented as the final result.

3.1.1 Operators

This section will detail the operators used within the algorithm implementation. The selection of operators aims to allow both exploitation and exploration of the search neighborhoods and offers a mix of faster and slower operators. A

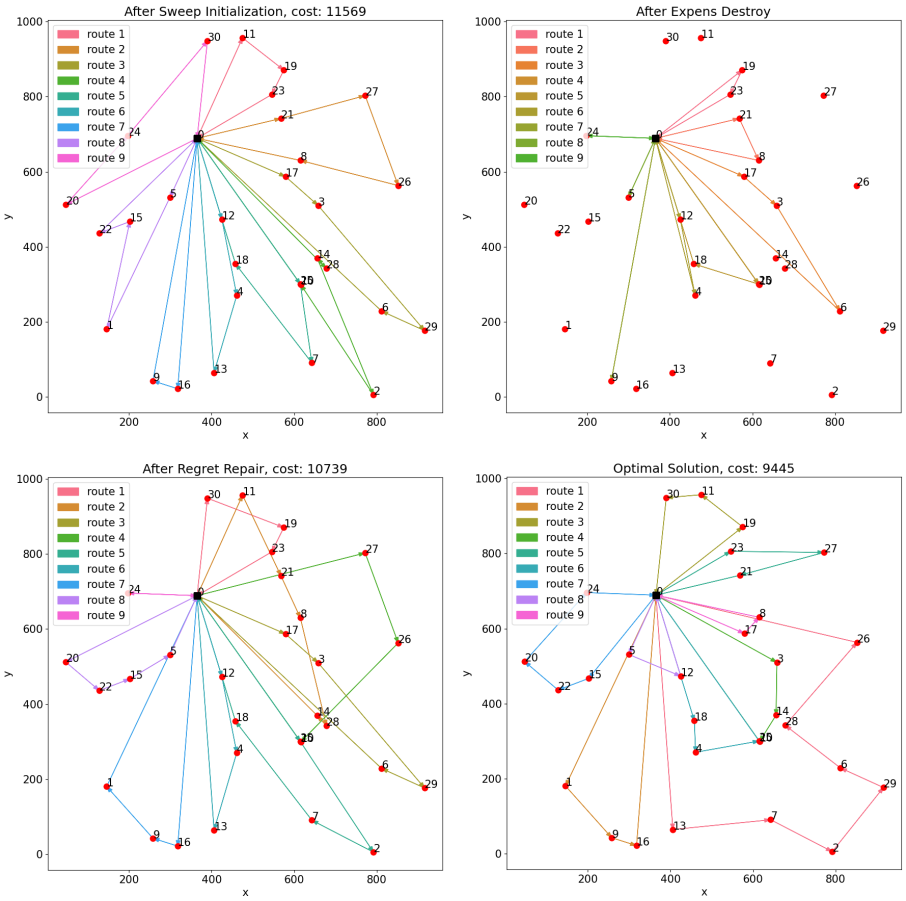

Figure 1: Four Example Solutions during an ALNS Run. Top left: initial solution, top right: destroyed solution, bottom left: repaired solution, bottom right: resulting optimal solution.

list of the most commonly used ALNS operators can be found in Windras Mara et al. (2022, p. 9).

Destroy Operators

The algorithm employs five destroy operators. Their individual degrees of destruction are governed by a meta-parameter detailed in chapter 4.4. Operators remove nodes until the desired degree of destruction is reached. The operators are as follows:

- **random_destroy** randomly removes nodes until the desired degree of destruction is reached.
- **nearby_destroy** randomly removes a single node. Additional nodes with the shortest edge distance to the initial node are removed.
- **demand_destroy** randomly removes a single node. Additional nodes with the lowest difference in demand compared to the initial node are removed.
- **expens_destroy** calculates the cost-savings that could be realized by removing a node. Nodes are removed in order of the highest costs.
- **routes_destroy** picks a random route and removes all nodes within it from the solution. It then tests if the desired degree of destruction is reached. If not, additional routes are removed. This operator can exceed the degree of destruction depending on the size of the last removed route.

Repair Operators

The repair operators are passed a list of removed nodes. The list is in order of removal. The repair operators will try to reinsert the removed nodes into feasible routes until all nodes are back in the solution. If no feasible route can be found, a new route is created. The operators are as follows:

- **greedy_repair** shuffles the list of removed nodes. Then, it finds the cheapest feasible route and position for the first node and inserts it into the solution. This is repeated until all nodes are repaired.
- **nearby_repair** takes the first node from the list of removed nodes. It then searches for the closest node (by edge distance) within a route feasible for insertion. The best position within this route is calculated, and the node is inserted. This is repeated for all removed nodes.
- **locost_repair** calculates the insertion cost and best position in each route for each node. It then places the node with the lowest insertion cost. After each insertion, the costs for inserting into the recently changed

route are recalculated, and all remaining nodes are resorted.

- **sorted_repair** runs the same initial sorting procedure as *locost_repair*. However, no recalculations are done. Nodes are inserted in the initial order and only check if their preferred route still has enough capacity. If not, the feasibility test is repeated until a valid route is found or a new route is created. While this operator was intended to be a faster, less precise version of *locost_repair*, testing shows that it often performs slower due to nested validity checks (see Figure 4).
- **regret_repair** calculates the cheapest and second-cheapest feasible insertion route for each removed node in the current solution. The node with the highest regret value, calculated as the difference between the second-cheapest and cheapest route, is inserted first. Insertion costs are recalculated after each placement, taking the recently placed node into account. This is repeated until all nodes are placed. The goal of *regret_repair* is to first insert nodes that would incur significant cost increases if their optimal route was filled up with other nodes.
- **regrdk_repair** spells out as Regret-Random-k Repair. k specifies which route's insertion cost is chosen for calculating the regret cost. For example, if k=4, the regret value will be the forth-cheapest route insertion cost minus the cheapest route insertion cost. Compared to k=2 regret repair, it favors nodes that might have two relatively cheap routes but a steep price increase afterward. This operator randomly chooses k so that $2 \leq k \leq min(5, number\ of\ routes)$.
- **friend_repair** is a complex insertion algorithm that calls other repair operators. It is based on finding a close-to-optimal solution in small neighborhoods (which we call friend-groups). First, we calculate the insertion cost for every removed node into every existing route. For each node, the best three routes are stored. A matrix saves a friend-score for each node pair by calculating the overlap in their three cheapest routes. For example, node 1 has cheapest insertions in routes A, B, and C. Node 2 is cheapest in routes B, C, and D. The overlap is 2. The node with the highest combined friend-score with all other nodes is chosen as the first friend. The three nodes with the highest friend-scores with the center-node are added to the friend-group, and those four nodes in the friend-group are removed from the list of removed nodes. The process of creating friend-groups is then repeated two more times.
  Now, we repeatedly call *greedy_repair* on a temporary copy of the active solution, with each permutation of the order of nodes in a friend-group as the nodes-to-insert. Since each friend-group consists of four nodes, 24 permutations are tested for each friend-group. For this use case, *greedy_repair* is set to not shuffle the order of the nodes-to-insert. The solution cost after insertion is tracked for each permutation, and the cheapest permutation is used on the active solution. This is repeated for each friend group. To insert the remaining removed nodes not part of any friend-group, *regret_repair* is called.
  The meta-parameters present in *friend_repair* have a massive influence on the runtime and performance of the operator. The n_r routes used for the overlap friend-score influence how far away from the cheapest route friends are noticed. The number of friend-groups and the number of nodes in each friend-group directly influence the runtime. In the case of the n_g friend-groups, the increase is linear, with the sorting and *greedy_repair* phase having to be repeated. In the case of the n_f friends per group, the number of permutations is multiplied by N whenever an N-th friend is added. For this implementation, the parameters were set as $n_r = 3$, $n_g = max(3, n_{removed_nodes}//n_f)$, and $n_f = 4$.

After the repair operation, a route-internal local 2-opt operation is run on every route that changed within the iteration. Positions of nodes within the route are swapped until no more improvements can be found. As the repair operators create solutions with varying levels of local optimization, the time needed for the 2-opt is added to the runtime of the repair operator. See Da Costa et al. (2021) for an in-depth explanation of the 2-opt heuristic.

3.1.2 Acceptance Criterion

As discussed in depth in Santini et al. (2018), ALNS uses an acceptance criterion after each iteration to decide which solution to proceed with. The criterion chosen for this implementation is Simulated Annealing, developed by Kirkpatrick et al. (1983) and Černý (1985). A comparison to other acceptance criteria can be found in Santini et al. (2018). Since this component of the meta-heuristic stays the same for adaptive- and learned operator-selection, fine-tuning it was not of high priority. Formula (1) shows the probability p to accept a new solution given a difference $\Delta c = c_S - c_{S^a}$ between the cost of the new solution and the accepted solution.

$$p = \begin{cases} 1 & \text{if } \Delta c \leq 0 \\ e^{\frac{-\Delta c}{sat_i}} & \text{if } \Delta c > 0 \end{cases} \tag{1}$$

Since the meta-heuristic has to be able to solve instances with significantly different cost ranges, the temperature is initialized after running the sweep-initialization as well as one iteration for each repair operator. Following Ben-Ameur (2004) and Kirkpatrick et al. (1983), the difference between the best and worst solution of these initial runs is used as the initial temperature.

The temperature is lowered each iteration by multiplication with a cooling rate $sa_cr < 1$, which gets initialized via formula (2), dependent on the maximum iterations I_{max}, as well as two meta-parameters governing the target remaining temperature sa_tp (as a percentage of the starting temperature) in a target iteration sa_ti. This approach ensures a

similar temperature curve over instances with different costs and optimization runs with varying maximum iterations.

$$sa_cr = sa_tp^{\left(\frac{1}{I_{max} \times sa_ti}\right)} \tag{2}$$

3.1.3 Adaptive Layer

The centerpiece of ALNS is the *adaptive layer*. It consists of two elements: the operator selection and the weight adjustment.

The operator selection phase chooses the destroy- and repair operators for the upcoming iteration independently of each other. It is implemented as a roulette wheel selection, with each operator's current weight w_o governing the probability p_o to be chosen. Since the selections for repair- and destroy operators are independent, the probability for a given repair operator o^+ calculates with $p_{o^+} = \frac{w_{o^+}}{\sum_o^{\Omega^+} w_o}$.

Adjusting the operator weights can be done in a multitude of ways. Most implementations sampled by Turkeš et al. (2021) follow close to the original proposal by Pisinger and Ropke (2007). Weights are updated after a batch of iterations. During each batch b, operators accumulate a score $\Phi_{o,b}$ dependent on the outcomes of iterations they were active in. We discern between five different outcomes ρ dependent on the cost of the new-found solution and the SA temperature: First: better than the currently best-known solution. Second: better than the currently accepted solution. Third: worse than the currently accepted solution but accepted due to simulated annealing. Fourth: rejected by simulated annealing. Fifth: identical to a solution found in a previous iteration. Each outcome is assigned a reward value ϕ_ρ, which is added to the operator score. Furthermore, the amount of times an operator is chosen within the current batch is tracked as $\Theta_{o,b}$, and the runtime of the selected operator is stored as $t_{o,i}$. We store a hash of each visited solution to be able to compare whether a solution is new.

We initialize all operator weights with $w = 10$ to ensure exploration of the available operators. After each batch, a weight update is performed using Formula (3). It is guaranteed that operators can never entirely leave the roulette selection by increasing the batch reward to at least one (Gullhav et al., 2017). The time factor $tf_{o,b}$ is calculated by Formula (4) where $\varnothing T$ is the average runtime of all operators so far (separated between repair and destroy). ω is the Time-Sensitivity Factor governing how sensitive the weight update is to longer operator runtimes. The minimizing term ensures that operators can only be punished for extended runtimes. Otherwise, operators could increase their weight simply by being fast without finding any improving solutions. Compared to the literature review in section 2.2.2, our approach tracks closest to Gullhav et al. (2017). However, we only calculate and apply tf at the end of a batch instead of after each iteration. This is done to average the runtimes of each operator.

$$w_{o,b} = (1-\alpha) \times w_{o,b-1} + \alpha \times max(1, \frac{\Phi_{o,b} \times tf_{o,b}}{\Theta_{o,b}}) \tag{3}$$

$$tf_{o,b} = min(1, \frac{\Theta_{o,b} \times \varnothing T}{\sum_{i=1}^{\beta} t_{o,i} \times \omega}) \tag{4}$$

Runtime Approximation

Timing the duration of a program operation is non-trivial. Simply using a function akin to a stopwatch to time from the start to the end of an operation is subject to noise, for example, another program slowing down the hardware an algorithm is run on. This noise also makes running a program multiple times with the same seed for random operations impossible, as operators' weights will fluctuate due to slightly different measured runtimes. As shown in section 1, some authors use the CPU cycle time to calculate the operator runtimes (Gullhav et al., 2017; Laborie & Godard, 2007). However, this is not possible for the environment in which our algorithm is run (Python on Windows). Instead, we implement a proxy: the runtime approximation (RTA) score. For this, the algorithm is run multiple times in a profiling mode, giving average runtimes for operators (e.g., *random_destroy*) and sub-operators (e.g., testing the feasibility of a route). From there, an RTA score is assigned to each relevant (sub-)operation. Additionally, some inner loops within the operators are given RTA scores to keep approximations close over wide ranges of customer node amounts. The individual scores are added whenever any function with an RTA score is called and stored as $t_{o,i}$ at the end of the destroy- or repair phase. While the approximations are not 100% accurate, this does not influence the study results. Both the ALNS and the LLNS are only aware of the RTA scores, and the final evaluation criterion is also based on RTA. This approach also allows us to artificially alter operators' efficiency by changing their rta components: We do this in the case of *sorted_repair* to distinguish its efficiency further from the similar *locost_repair*. Figure 2 compares approximated and actual runtimes (in nanoseconds).

Figure 4 in Section 4.3 shows how measured and approximated runtimes behave with growing instance size.

3.2 Learned Large Neighborhood Search

This section explains the setup for training our LLNS policy. We introduce the supplied features, explain potential reward functions, and discuss important hyper-parameters.

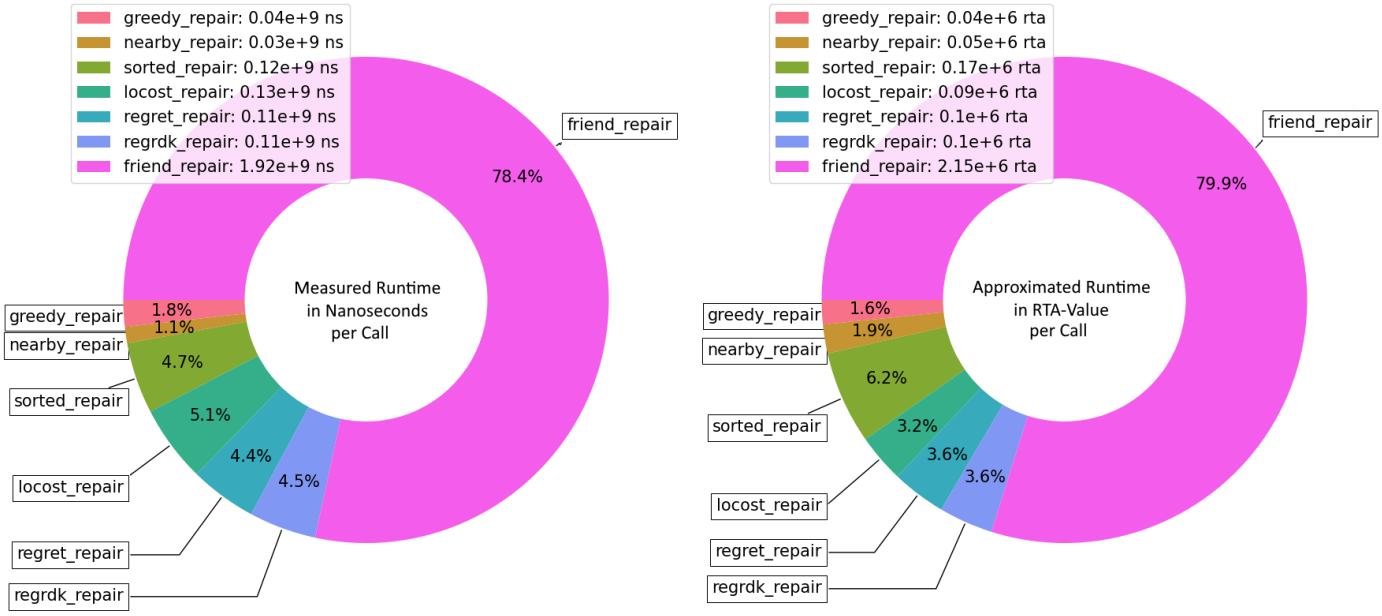

Figure 2: Operator Runtime Comparison: Average Measured Runtime in Nanoseconds and Calculated Approximated Runtime per Call. 90 (¬)ALNS Evaluation Runs, 143-351 Nodes.

3.2.1 Features

One advantage of a trained agent over the adaptive layer is that it can take in much more varied information. The adaptive layers' decisions are only informed by the previous performance of the operators within the current optimization run. An agent can not only utilize information from past runs but can also be fed more data about the current state of the optimization problem. Our goal for the feature selection is to prefer domain-agnostic information to be able to transfer the approach to different problem settings.

Initial testing was done on a small subset of observations, only including the four first features (*completion_percentage, destroy_operator, number_of_destroyed_nodes, iterations_since_last_best*) of the following list. As further discussed in the results in Section 4.5, this proves to be too little information to outperform a handpicked pre-selection of operators. Consequently, we add six additional features:

- **completion_percentage** is the ratio of the already completed iterations to the maximum number of iterations. It is presented as an integer between zero and 100.
- **destroy_operator** informs which destroy operator was used in the current iteration. Since the destroy operators might create sets of non-serviced nodes with different characteristics, some repair operators might perform better after a certain destroy operator was called. The agent has no information on the actual behavior of the destroy operators; it is only provided an index from the list of destroy operators.
- **number_of_destroyed_nodes** carries two insights: first, it can enable choosing repair operators that perform better with more or less destroyed nodes. The RTA of the repair operators is highly sensitive to this number since some operator's RTA scales linearly (e.g., *greedy_repair*), whereas other operators need to repeat calculations for all remaining nodes whenever a node is inserted (e.g., *regret_repair*). The second insight is information on the number of nodes in the problem setting. Since destroy operators have fixed degrees of destruction, knowing both the destroy operator and the number of destroyed nodes tells the agent the size of the problem. As the agent is trained on instances between 100 and 399 nodes, it is possible to learn which operators perform better in smaller or larger instances. Because the training set (see section 4.1) includes instances with a distinct number of nodes, and we do not want the agent to over-train to a point where it optimizes based on the characteristics of the exact instance, a random noise-term of [-1, 0, +1] is added to this observation.
- **iterations_since_last_best** provides information on how long the algorithm has been stuck without finding a new best solution. It ignores solutions accepted only by simulated annealing. Its goal is to enable the agent to choose operators that facilitate either exploration (after no improvements have been found for a long time) or exploration (of new-found improvements). This observation is presented as a percentage of the maximum iterations.
- **current_rta_sum** informs the agent of how much time has been spent in the run so far. As it is a direct part of the reward function (7), this observation is provided to enable the agent to better estimate expected rewards.
- **solution_change_best_vs_initial** is used as a proxy for

the second part of the reward function: the GTO. As we cannot expect to know an optimal solution for any problems outside the training set, we inform the agent of how much it has managed to improve the initial solution. Because of diminishing returns in the search process, the growth of this observation should slow at a similar rate as the convergence of the solution towards the optimum.

- **rta_last_iteration** informs the agent about the duration of the last repair operation.
- **solution_change_last_iteration** informs the agent about the change in objective value within the last iteration. It is calculated as $\frac{c_S}{c_{S^a}}$ where c is the objective value.
- **acceptance_check_result** presents a value between zero and four, corresponding to the SA outcome (new_best / accepted / annealed / rejected / already_visited) of the last iteration.
- **time_complexity_last_operator** groups the repair operators according to their runtime behavior. Figure 4 shows operator runtimes in relation to instance size. Group 1: *greedy_, nearby_*. Group 2: *regret_, regrdk_, locost_*. Group 3: *sorted_*. Group 4: *friend_repair*. This feature is intended to speed up the learning progress in regards to the operator runtime behavior.

3.2.2 Reward Function

Table 3 shows that different reward functions are feasible for LLNS. We identify three approaches: (Reijnen et al., 2022) and Boualamia et al. (2023) give a reward in each iteration with the score ϕ_ρ dependent on the outcome ρ of the acceptance check. This copies the weight adjustments from ALNS. (Bongiovanni et al., 2022) give a reward in each iteration dependent on the objective value change within the last iteration. (Johnn et al., 2023) reward only once, in the last iteration, dependent on the objective value change between the initial and final solution.

We test approaches similar to the three mentioned and find arguments for and against each. As we maximize rewards, reward values for time spent are negative. We use a discount rate of 1, which means future rewards are fully calculated into every previous step within a run:

- Copying the **ALNS reward scheme** as shown in Formula (5) requires no additional work if ALNS is already implemented. Furthermore, compared to the following two approaches, it does not require knowledge of a lower bound solution value for the training dataset. However, it does not distinguish between small and big improvements. A well-trained policy might optimize rewards gained by finding as many small improvements as possible instead of searching for the best final result.

$$R_i = \phi_{\rho_i} \tag{5}$$

- Giving a **single reward in the final iteration** as shown in Formula (6) is straightforward and enables lining up the reward with the algorithm comparison criteria, in our case the expected runtime approximation value at 5% gap to optimum (eRTA-GTO5) explained in Section 4.2. The reward value is the negative cumulative RTA whenever eRTA-GTO5 was first reached, or the expected eRTA-GTO5 if 5% gap to optimum (GTO5) was not reached. This approach has the downside of also rewarding any actions after GTO5 is reached. The policy might make suboptimal decisions after this point, and these actions - which had no impact on the final result - will gain the same rewards as the relevant actions before reaching the GTO-target. This is a downside to our decision to run fixed iterations instead of a fixed runtime (compare Section 2.2.1) and could be mitigated by running for a fixed time and then rewarding based on the reached solution quality.

$$R_i = \begin{cases} 0 & \text{if } i \neq I_{max} \\ -1 \times \text{eRTA-GTO5} & \text{if } i = I_{max} \end{cases} \tag{6}$$

- As we want to optimize for both runtime and solution quality, giving a **reward in each iteration** forces us to consider both the current runtime and the current solution quality. Therefore, we need a formula that can balance the trade-off between runtime and solution quality. We manage this by multiplying the current cumulative RTA with the currently reached GTO+1 $= (\frac{c_{S^*}}{c_{S^{opt.}}})$. As both sides of the multiplication scale differently, we need a balancing factor, which we call the gap-factor gf. This factor must be tuned to achieve the correct trade-off between fast runtimes and good solution quality (see Section 4.4). This complicated approach has the upside of correctly rewarding actions even after the desired GTO-target is reached. Formula (7) shows our reward function for rewards R in each iteration.

$$R_i = -1 \times (\frac{c_{S^*}}{c_{S^{opt.}}})^{gf} \times rta_sum_i \tag{7}$$

If no lower bounds are known for the training data set, the objective value improvement between the initial and rewarded solution can be used as a proxy for the GTO (Johnn et al., 2023). A short discussion of the performance of the three reward schemes can be found in 4.5.

3.2.3 Training Setup

The training setup for the LLNS agent is built to rely on publicly available libraries. We train a Deep-Q Network (DQN) using TensorFlow's *tf_agents* package (Developers, 2023). We test both the standard DQN-agent and the C51-agent, which calculates a Q-value-distribution instead of a single Q-value (Bellemare et al., 2017). As the outcome of each iteration is heavily stochastic, the distribution might aid the training, but we find no significant differences between DQN- and C51-agents for our application. We implement a replay buffer with a size of 10% of all training steps using the *tf_uniform_replay_buffer* package (Developers, 2023). Extensive testing is needed to find a learning

rate that neither gets stuck in a local optimum nor cannot keep any learned information. We find that a decaying learning rate works best. Special attention is needed for the epsilon-greedy factor ϵ, which determines how often a random action should be taken during training. As every step influences both the solution quality and the runtime, too many random actions can prohibit the learning of strong policies. We choose a decaying ϵ.

We build our ANN with three hidden layers of size [256, 128, 64], following the theoretical advice by Stathakis (2009) and Bengio (2012), as well as the LLNS-implementation by Boualamia et al. (2023). Training runs featured up to two million transitions, calculated live using a set of training instances with between 101 and 393 nodes. This leads to training runs between 24 and 92 hours. Since the majority of training time is spent solving the routing problem, we believe that significant time could be saved by switching to learning in an offline transition database.

Section 4.5 explains our learnings from the training runs. Appendix Table A3 lists all hyper-parameter values.

4 Results

The following Chapter presents our results for analyzing the performance of our proposed extensions to the ALNS. We start by outlining the comparison setting by explaining the instance datasets (Section 4.1), the multi-objective comparison criterion (Section 4.2), and by presenting a quantitative analysis of the repair operator performance (Section 4.3).

We then present the results of our meta-parameter tuning (Section 4.4) and report on the behavior of the LLNS training runs with varying hyper-parameters, feature sets, and reward functions (Section 4.5).

Section 4.6.1 presents the main results of our study, the comparison of ALNS, (¬)ALNS, tALNS and LLNS. Subsections 4.6.2 and 4.6.3 present additional results for the behavior of tALNS with different operator portfolios, as well as a test of the generalization capabilities of tALNS and LLNS in larger instances (700+ nodes).

To reduce variability, we only use *random_repair* with $dod = 0.33$ for our training- and evaluation-runs. Testing for impact on the eRTA-GTO5 of the tALNS implementation, we found only a non-significant change (-1.37% average eRTA-GTO5, p-value 0.888) compared to using all destroy operators.

4.1 Instance Dataset

The publicly available CVRP-library by Uchoa et al. (2017) is used to generate problem instances. It is split into a training set and an evaluation set, with the evaluation set containing three files for each century of customer nodes. In this context, *century* refers to a group of instances with the same hundredths place value in their number of nodes, e.g., 100 - 199 nodes are grouped into century 100. To test the algorithms under varying conditions, we select test instances with varying characteristics of depot position, customer clustering, and minimum vehicles needed. Appendix Table A1 gives detailed information on the evaluation instances. A plot for each instance can be found on the CVRPLIB website (Lima et al., 2023). The reward functions for LLNS training and the algorithm comparison criterion rely on a gap to optimum (GTO) value. (Uchoa et al., 2017) provide the best-known solution with each of their problem instances, which, for most instances, is the optimum solution proven by exact solving. The best-known solution is used as a stand-in if the optimal solution has not been verified.

4.2 Algorithm Comparison Criterion: Runtime at Fixed Solution Quality

As shown in section 2.2.3, a straightforward method to judge different meta-heuristics for both speed and performance is to compare the solution quality after a fixed runtime is reached. Due to multiple reasons, we decide against this approach: Firstly, SA lowers the temperature after each iteration. Therefore, not using fixed iteration targets leads to unpredictable temperature behavior. Secondly, fixing iterations allows for a degree of automatic adaption of runtimes to problem size. As the degree of destruction of each destroy-operator is a fixed percentage of the instance nodes, each iteration's neighborhood search grows in size when more nodes are considered. These larger sub-problems take more time to solve, leading to overall longer solving times for larger instances. This creates a need to consider different runtime-stopping-criterions dependent on instance size, increasing complexity. Section 4.3 explores the runtime of repair-operators in relation to instance size. Lastly, within training, the agent is rewarded after each iteration, with the outcomes of later iterations influencing early decision-making via discounted rewards. An unpredictable number of future rewards due to fluctuating iterations adds an additional undesirable variable to the training process. For these reasons, we decide to keep the fixed iterations also used by Pisinger and Ropke (2007).

In the sampled literature, none of the authors using an iteration-based stopping criterion implements a quantitative approach to judge algorithms on speed and solution quality. Instead, both measures are reported and compared independently (Wu et al., 2017; Zhang & Yang, 2021). Gullhav et al. (2017) stop after fixed iterations but judge solution quality at three fixed durations. This is not feasible for our approach, as the diverse problem instances and operator runtimes lead to massive differences in solving times. For example, our ALNS capped at 2000 iterations can solve a simple layout with 150 customers to optimality after 30 seconds but can take up to 30 minutes to reach a 5% GTO in a very large instance with small transport capacity per vehicle.

A lower outcome variability is observed for the final GTOs. In 180 runs of ALNS in the evaluation dataset, no run exceeds 9% GTO, and the standard deviation of GTOs is measured at 0.022 (see Table 7). Therefore, we decide to use the RTA needed to get to GTO5 as the comparison criterion. For runs that do not reach GTO5, an expected runtime

approximation value at 5% gap to optimum (eRTA-GTO5) is calculated. Algorithm 2 shows how the expected runtime approximation (eRTA) is calculated using the expectation factors shown in Table 4. The eRTA factors are derived by running 300 instances for 10,000 iterations each. The ratios between the median RTAs at each relevant GTO-milestone are calculated for centuries 100, 200, and 300. Instances with more than 399 nodes use the factors of century 300.

Runs that do not reach GTO10 are considered as *failed* and are punished with the highest eRTA-GTO5 out of any compared algorithm runs on the given problem instance. Algorithm results that include runs which do not reach GTO10 will always be clearly marked as such. All runs within the main algorithm comparison in Table 7 reach GTO10.

Table 4: Expectation Factors for Runtimes at Gap to Optimum Values

Century	0.10 → 0.09	0.09 → 0.08	0.08 → 0.07	0.07 → 0.06	0.06 → 0.05
100	1.63	1.34	1.78	1.38	1.37
200	1.62	1.87	1.72	1.60	1.24
300	2.25	1.87	1.75	1.26	1.48

Century refers to the instance size, with all instances between 100 and 199 nodes being sorted into century 100.

Cell values display the factors for estimating the time needed to reach a smaller gap to optimum.

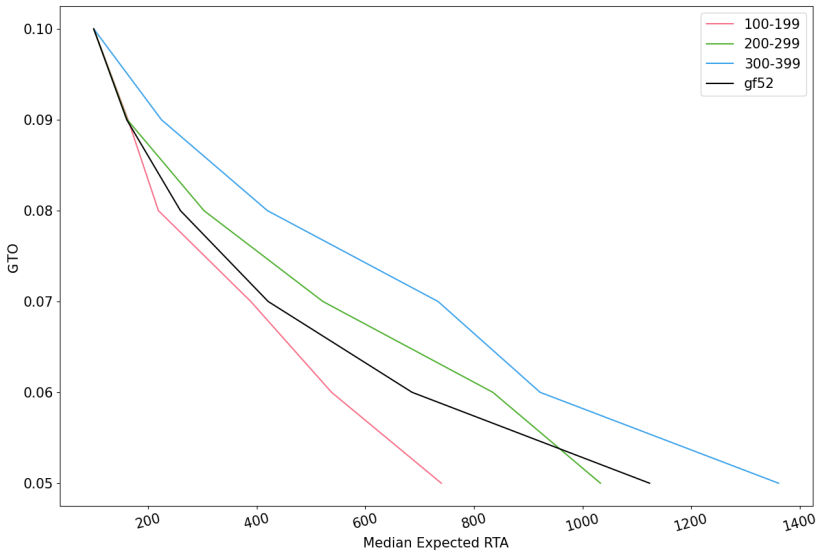

Figure 3: Colored: Curves for the Expected RTA, starting at RTA-GTO10=100, by Instance Size Century. Black: Curve for the Gap Factor gf=52.

Algorithm 2: Algorithm to Calculate eRTA for a given Century

```
Inputs: observed_rta_at_gto, eRTA_gto_factors[century]
rta_steps = [0.09, 0.08, 0.07, 0.06, 0.05]
if observed_rta_at_gto[0.10] == 'not_reached' then
    return '10%_not_reached'
else
    eRTA_at_gto[0.10] = observed_rta_at_gto[0.10]
for gto in rta_steps do
    if observed_rta_at_gto[gto] != 'not_reached' then
        eRTA_at_gto[gto] = observed_rta_at_gto[gto]
    else
        eRTA_at_gto[gto] = eRTA_at_gto[gto + 0.01] *
          eRTA_gto_factors[gto + 0.01]
return eRTA_at_gto
```

4.3 Repair Operator Comparison

Figure 4 shows repair operator runtime behavior in growing instances. We observe that the RTA-approach does not accurately model the growing runtimes in larger instances. Furthermore, measured nano-second (ns) runtimes peak in instances with the depot in a corner, while the calculated RTA spikes with higher vehicle counts. This points towards skewed RTA-values for some suboperations. Section 3.1.3 explains why this does not invalidate our algorithm comparisons. See Appendix Table A1 for an overview of instance layouts. Appendix Table A2 shows mean and median operator runtimes for all instances up to 600 nodes.

Nevertheless, in both ns- and RTA-runtimes, it is clear that operator runtimes differ significantly, with an average call of *friend_repair* taking more than 50 times longer than *nearby_repair*. The classic operators like *greedy_repair* and *regret_repair*, which are routinely used in ALNS-literature,

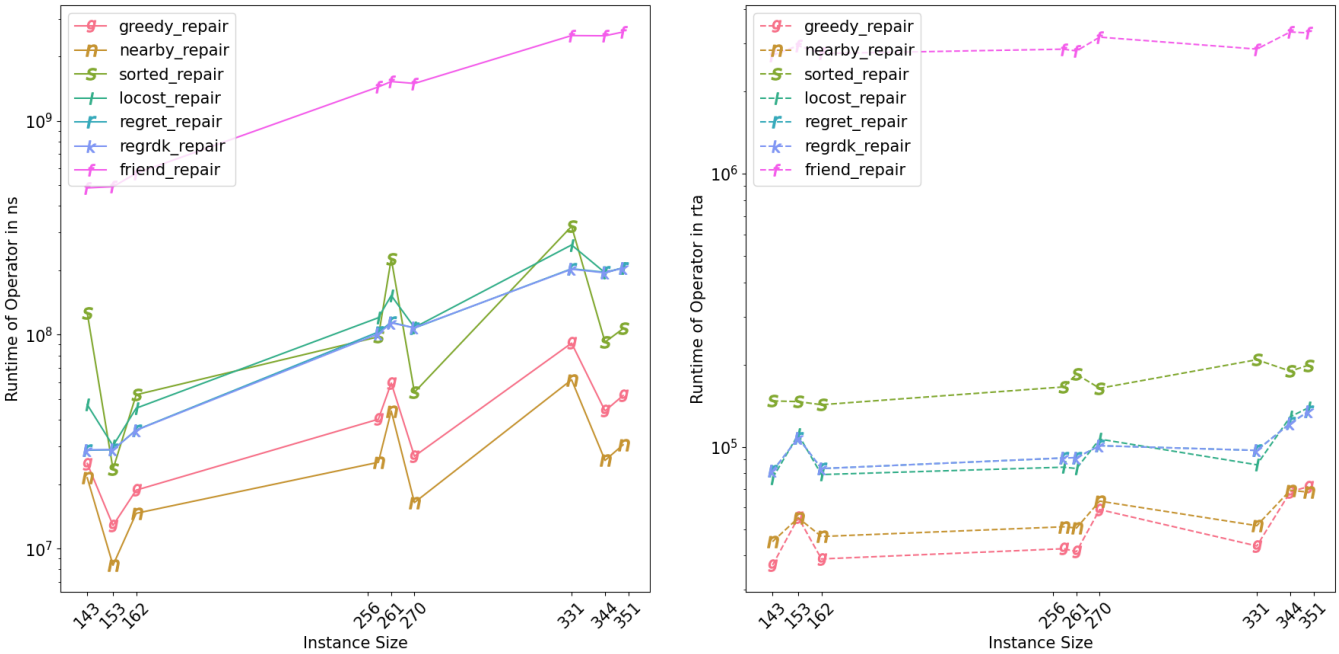

Figure 4: Operator Runtime in Observed Nanoseconds (left) and Calculated RTA (right). 10 (¬)ALNS runs per instance. Logarithmic y-scale.

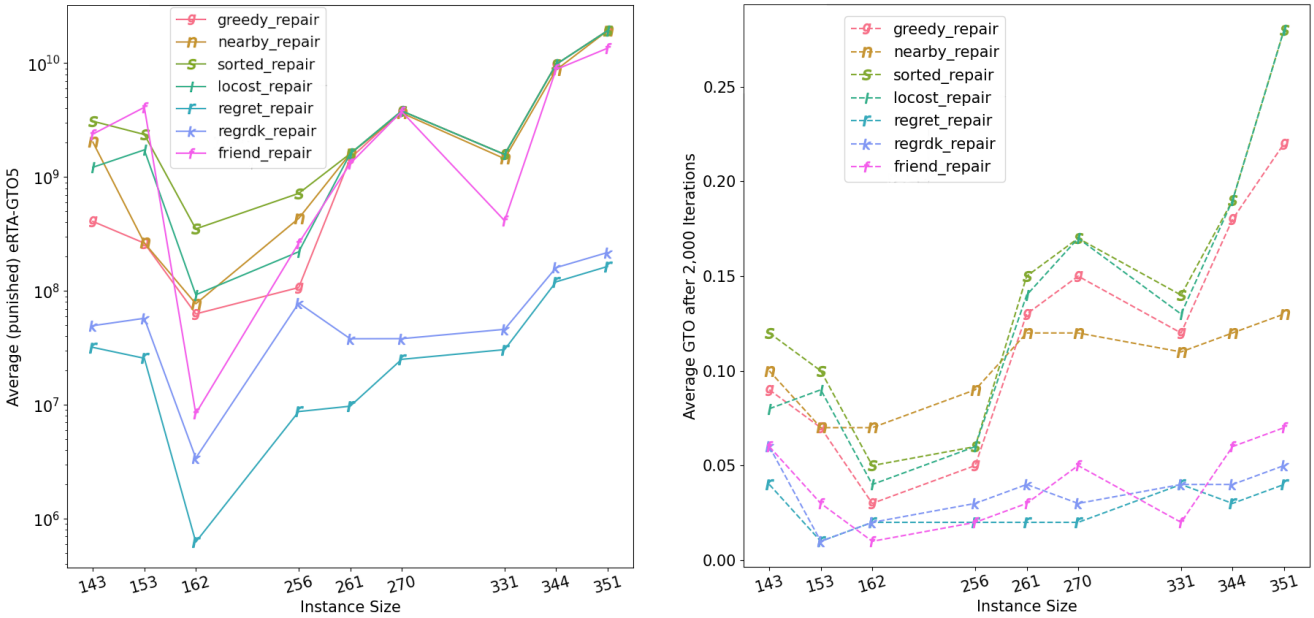

Figure 5: eRTA-GTO5 (left) and GTO after 2000 Iterations (right) reached by Runs with Single Operator Portfolios. Logarithmic y-scale for left plot. 10 runs per instance and operator.

still differ by a factor of 10 (Windras Mara et al., 2022). Profiling runs show that for our algorithm implementation, repair operators take up between 75% and 95% of the metaheuristic's runtime, the remainder being split between the destroy phase and miscellaneous operations like weight adjustments or acceptance checks. These timing results clearly motivate the need for a time-sensitive operator weight adjustment scheme.

Figure 5 displays operator performance in Single Operator Portfolio runs. This mimics testing a LNS for each operator. The findings show that *regret_repair* has the highest efficiency, *friend_repair* finds good solutions but is comparatively slow, and the fastest four operators are unable to reach GTO10 in larger instances.

In addition to the LNS-results, we analyze operator performance in (¬)ALNS. These results are less conclusive and

Table 5: Meta-Parameter Tuning Values

Parameter	(¬)ALNS	ALNS	tALNS	LLNS	Range	Value
ϕ_ρ	X	✓	✓	X	[20-100, 5-50, 5, 1, 1]	[50, 10, 5, 1, 1]
ω	X	X	✓	X	0.1 - 1.5	1.20 (all operators) 0.15 (pre-selected*)
β	X	✓	✓	X	10 - 100	50
α	X	✓	✓	X	0.1 - 0.9	0.5
dod	✓	✓	✓	✓	0.1 - 0.5	0.33
gf	X	X	X	✓	not applicable	52

ϕ_ρ: Acceptance Reward for Result ρ. ω: Time Sensitivity Factor. β: Batch Size. α: Weight Update Rate.
dod: Degree of Destruction for Destroy Operators. gf: Gap Factor for LLNS-rewards.
***pre-selected** refers to a reduced operator portfolio of *greedy_, nearby_, locost_* and *random_repair*

can be found in Appendix Section A5.

4.4 Meta-Parameter Tuning

ALNS contains several parameters that need tuning for the algorithm to perform optimally (Ropke & Pisinger, 2006). We will refer to them as meta-parameters to distinguish them from parameters that influence Reinforcement Learning (RL), often called hyper-parameters. As the objective of this study is to compare different versions (ALNS, (¬)ALNS, tALNS, LLNS) of the same meta-heuristic implementation, we reduce complexity by only tuning meta-parameters which influence the performance of a subset of the compared algorithms. Therefore, we leave, for example, SA-parameters untuned, as they do not impact the comparison of our ALNS-variants. We use starting values from the existing literature for these non-tuned parameters and make slight adjustments by hand. Appendix Table A3 gives a full list of all final parameter values.

As meta-parameter values influence each other's best value, evaluations must be conducted for each combination of values in the testing ranges. We reduce complexity by forming smaller clusters of meta-parameters. The clusters are then optimized in sequence. We group the acceptance check rewards ϕ with the time-sensitivity factor ω as both directly affect the weight update values. Furthermore, we group the batch size β with the weight update rate α, as their combination controls how fast weights can fluctuate. Table 5 lists the tuned parameters and which algorithm variants they affect.

A challenge of tuning the meta-parameters is dealing with different instance sizes: Tuning is done on instances between 101 and 393 nodes (the training set for LLNS training). The optimized objective value is the eRTA-GTO5. Since this value scales with instance size and layout, the standard deviation of tuning results over a wide span of instances tends to be very high. Therefore, reaching a tuning result with low p-values requires excessive runs. Filtering by instance century reveals that different meta-parameter values excel for different centuries (and most likely also other problem characteristics like the instance layout). P-values reach significantly smaller values (<0.05) when analyzing century averages instead of averages over all tuning instances. Nevertheless, a single meta-parameter value has to be chosen for all centuries. If different values are optimal for different instance size centuries, we opt for the best value for century 200.

For the LLNS, we only need to tune one meta-parameter, the gap factor gf. The hyper-parameter selection for RL training is explained in Section 4.5. The gap factor determines how expired runtime and reached GTO are weighted against each other in Formula (7). It balances the trade-off between fast runtimes and good solution quality. We already calculate the expected runtime behavior in relation to GTO-reached for the eRTA-factors. We determine the gf by matching it to the eRTA-curve for century 200 using Formula (8). Figure 3 shows the eRTA-curves and a curve for $gf = 53$.

In Formula (8), $eRTAf_{0.10\rightarrow gtd}$ refers to the compounding eRTA-factors (of century 200) between GTO10 and the GTO-milestones we are summing over. Table 4 lists milestones and eRTA-factors. We chose $gf \varepsilon \mathbb{N}^+$ so that the squared error between the GTO- and the RTA-parts of the formula is minimized. This ensures that the Reward Formula (7) produces similar rewards for any points (RTA/GTO) on the eRTA-curve for century 200.

$$minimize \sum_{gto=0.05}^{0.10} \left(\frac{1.1^{gf}}{(1+gto)^{gf}} - eRTAf_{0.10\rightarrow gto}\right)^2 \quad (8)$$

4.5 Q-Learning Training Results

As stated in Section 3.2.3, we test multiple different learning rates, ϵ-greedy values, feature portfolios, and reward functions for our LLNS-training setup. A statistical analysis of the performance of given combinations of hyper-parameters and environment setups is out of the scope of this study. This section will nevertheless give a short review of our training results.

After tuning by hand, we settle on a decaying ϵ for ϵ-greedy, starting at 80% and lowering over the first 30% of episodes to 0.1%. The learning rate starts at $1e^{-}4$ and decays to $1e^{-}6$ over the first 75% of episodes.

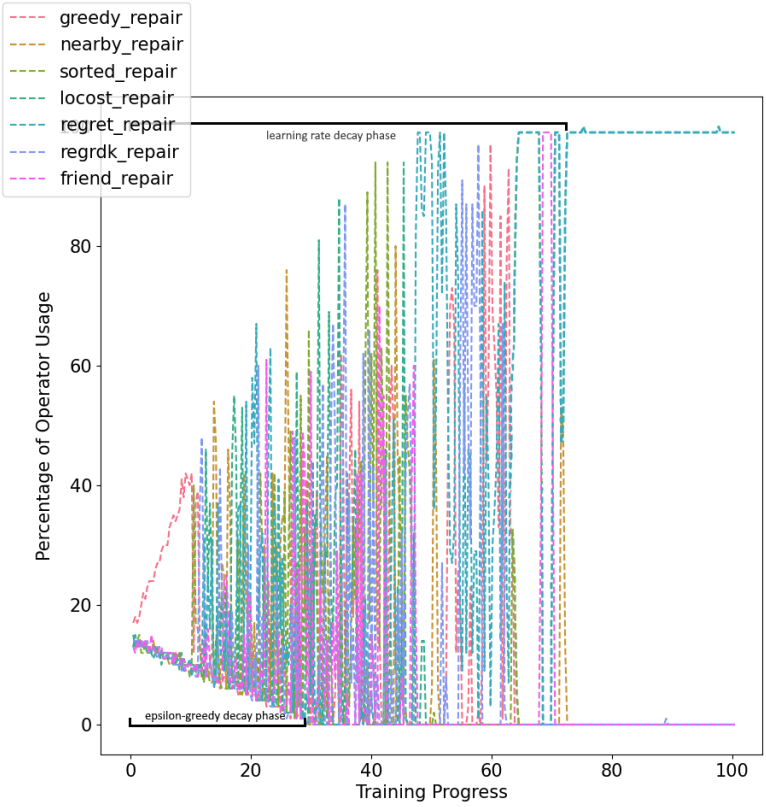

Figure 6: Operator Selection Percentage during Training. Section 0-30 shows decaying ϵ. Section 75-100 shows fully decayed Learning Rate.

Figure 6 shows the policy behavior during a training run. The decaying ϵ-greedy is visible between 0% and 30% training progress. After 75% progress, we observe fewer changes in the policy, which is caused by the decayed learning rate. The policy converges towards choosing *regret_repair* in almost all iterations.

Out of the three tested reward schemes discussed in Section 3.2.2, we do not converge to a strong policy within a reasonable training time using ALNS-rewards. Both the multi-objective rewards in each iteration as well as the single eRTA-GTO5-reward in the final iteration reach policies relying solely on *regret_repair*, with the single final reward scheme doing so with slightly higher consistency.

Furthermore, we test two different sets of features. The small set (first four features mentioned in 3.2.1) does not converge to *regret_repair* without decaying ϵ-greedy and learning rate. The large feature portfolio does so, but only after long training times. With the decaying hyper-parameters implemented, both feature portfolios converge to using only *regret_repair*.

Whether the trained selection policies convergence towards single-operator usage is caused by the overwhelming effectiveness of the chosen operator (potentially exacerbated by the RTA-calculations of the operators) or by insufficient training is subject to further research. For now, we can show that LLNS is able to select the most efficient operator, effectively turning ALNS back into LNS, but without the need for hand-picking an operator. We cannot prove that no better policy utilizing more repair operators exists. We test the training setup with *regret_repair* and *regrdk_repair* removed, but are unable to learn strong policies for this setup. Further research is needed to investigate whether this issue is caused by unfit hyper-parameters for the smaller operator portfolio, insufficient training times, flaws in our general training setup, or other unknown causes.

4.6 Algorithm Comparison

This section presents and discusses our main algorithm comparison results. Furthermore, we present additional tests for tALNS-behavior under different operator portfolios as well as the generalization capabilities of our algorithms.

4.6.1 Algorithm Comparison Results

Figure 7 shows which repair operators are preferred by different algorithms. Two operator portfolios are present: The full portfolio of all seven repair operators, and a pre-selected (PS) portfolio that simulates an algorithm designer hand-picking promising operators for a given problem domain. We can observe how the average eRTA-GTO5 decreases with declining usage of *friend_repair* and increasing usage of *regret_repair*. We can also see the impact of a

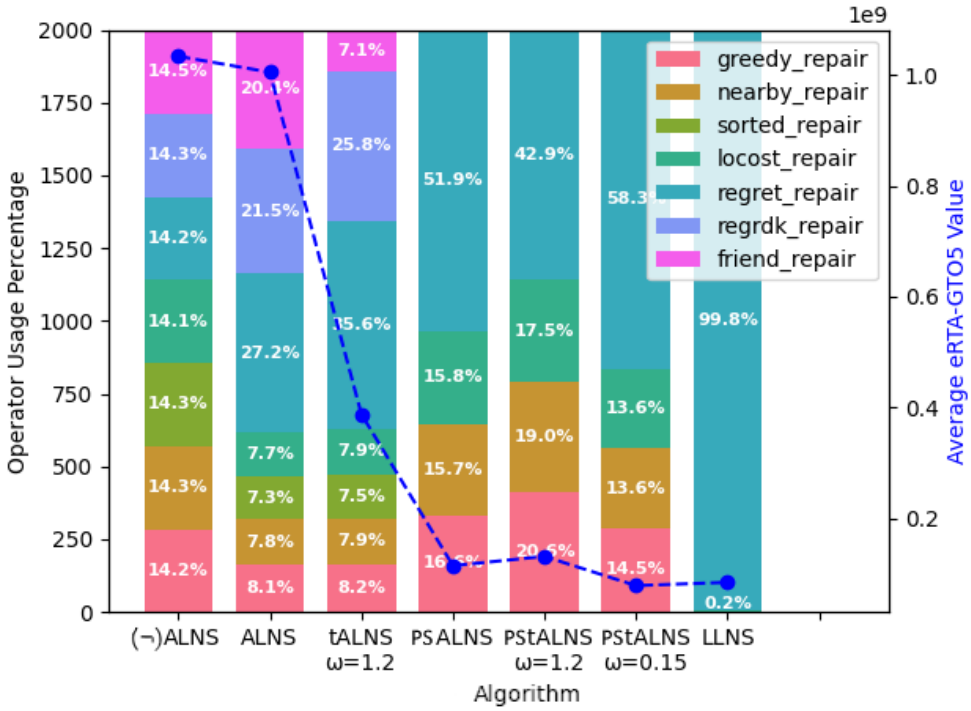

Figure 7: Operator Selection Percentage of Compared Algorithms as Stacked Bars. Overlaid is the average eRTA-GTO5 of the algorithms in blue. LLNS chooses *locost_repair* in 0.2% of iterations.
ps refers to operator pre-selection: *greedy_*, *nearby_*, *locost_* and *random_repair*.
ω is the Time-Sensitivity Factor.

Table 6: Algorithm Performance as Matrix Comparison

	(¬)ALNS	ALNS	tALNS, $\omega = 1.2$	PS ALNS	PS tALNS, $\omega = 1.2$	hp tALNS, $\omega = 0.15$
ALNS	-2.77%, p=0.7026					
tALNS, $\omega = 1.2$	-62.49%, p=0.0000	-62.49%, p=0.0000				
PS ALNS	-88.94%, p=0.0000	-88.94%, p=0.0000	-70.50%, p=0.0000			
PS tALNS, $\omega = 1.2$	-87.32%, p=0.0000	-87.32%, p=0.0000	-66.19%, p=0.0000	+14.61%, p=0.0326		
PS tALNS, $\omega = 0.15$	-92.42%, p=0.0000	-92.20%, p=0.0000	-79.89%, p=0.0000	-31.48%, p=0.0000	-40.22%, p=0.0000	
LLNS	-91.85%, p=0.0000	-91.85%, p=0.0000	-78.27%, p=0.0000	-26.32%, p=0.0079	-35.71%, p=0.0001	+7.53%, p=0.5954

Algorithm in row compared to Algorithm in column. Percentage values calculated as average eRTA-GTO5 of $\frac{row}{column} - 1$.
p-values are two-sided. ω is the Time Sensitivity Factor. **PS** refers to operator pre-selection.

correctly tuned sensitivity factor ω for the time-sensitive ALNS, depending on the operator portfolio. This topic is further explained in Section 4.6.2.

As the differences between our proposed algorithms and classic ALNS shown in Table 7 are too large to provide meaningful p-values, we display a direct comparison between each algorithm in Table 6.

Table 7 shows the average eRTA-GTO5 as the main algorithm comparison criterion. As the trade-off between fast runtimes and good solution quality might change depending on problem scenarios, we also report the final GTO and RTA after 2000 iterations. Lastly, the average iteration of the best solution is reported. This value allows us to estimate whether adding additional iterations would lead to better solution quality, i.e., it tells us if there is room for improvement left at the end of the 2000 iterations.

Table 7: Results of Algorithm Performance Comparison

Instance	Avg. eRTA-GTO5	Std. Dev. eRTA-GTO5	Avg. Final GTO	Avg. Final RTA	Avg. Iter. of Best Solution	% Change to ALNS	T-Value	p-Value***
				(¬)ALNS				
143-k7-corner	4.96E+8	1.10E+8	2.50%	9.12E+8	1857	-1.67%	-0.24	0.8143
153-k22-clustr	6.81E+8	1.16E+8	2.60%	10.18E+8	1861	+20.92%	3.20	0.0108
162-k11-center	2.04E+8	1.82E+8	0.90%	9.39E+8	1779	+128.42%	1.99	0.0773
256-k16-clustr	3.21E+8	1.13E+8	1.60%	9.57E+8	1735	+33.06%	2.24	0.0521
261-k13-corner	6.14E+8	1.53E+8	3.30%	9.90E+8	1903	-8.58%	-1.19	0.2653
270-k35-center	11.91E+8	3.71E+8	5.20%	10.85E+8	1853	+9.97%	0.92	0.3814
331-k15-corner	5.85E+8	1.53E+8	3.60%	10.07E+8	1659	-22.97%	-3.59	0.0058
344-k43-center	20.65E+8	5.92E+8	6.70%	11.42E+8	1746	-7.31%	-0.87	0.4066
351-k40-clustr	31.46E+8	9.47E+8	7.40%	11.57E+8	1788	+8.29%	0.80	0.4420
Average*	10.34E+8	9.99E+8	3.76%	10.23E+8	1798	+2.85%	0.27	0.7864
				ALNS				
143-k7-corner	5.05E+8	2.92E+8	2.25%	13.63E+8	1755			
153-k22-clustr	5.63E+8	1.63E+8	1.80%	10.21E+8	1752			
162-k11-center	0.89E+8	0.87E+8	0.50%	13.12E+8	1674			
256-k16-clustr	2.42E+8	1.17E+8	1.25%	15.28E+8	1672			
261-k13-corner	6.72E+8	2.70E+8	2.60%	13.14E+8	1763			
270-k35-center	10.83E+8	2.88E+8	4.50%	13.42E+8	1793			
331-k15-corner	7.59E+8	3.69E+8	3.30%	16.13E+8	1732			
344-k43-center	22.28E+8	8.50E+8	5.95%	14.45E+8	1824			
351-k40-clustr	29.05E+8	9.12E+8	6.85%	14.84E+8	1854			
Average**	10.05E+8	10.05E+8	3.22%	13.80E+8	1758			
				tALNS, ω=1.2				
143-k7-corner	1.46E+8	0.60E+8	1.10%	5.46E+8	1585	-71.15%	-18.87	0.0000
153-k22-clustr	2.54E+8	0.61E+8	1.50%	5.40E+8	1904	-54.86%	-16.11	0.0000
162-k11-center	0.56E+8	0.48E+8	0.90%	4.97E+8	1638	-37.35%	-2.20	0.0557
256-k16-clustr	0.82E+8	0.41E+8	1.20%	5.36E+8	1721	-66.05%	-12.22	0.0000
261-k13-corner	1.85E+8	1.11E+8	2.40%	6.36E+8	1481	-72.49%	-13.93	0.0000
270-k35-center	5.96E+8	1.10E+8	4.80%	7.07E+8	1833	-44.98%	-14.03	0.0000
331-k15-corner	2.86E+8	0.72E+8	3.20%	5.79E+8	1784	-62.38%	-20.66	0.0000
344-k43-center	8.67E+8	2.93E+8	6.00%	7.21E+8	1785	-61.08%	-14.68	0.0000
351-k40-clustr	10.18E+8	2.25E+8	7.00%	6.92E+8	1763	-64.96%	-26.52	0.0000
Average*	3.88E+8	3.61E+8	3.12%	6.06E+8	1722	-61.43%	-16.21	0.0000
				Pre-Selected ALNS****				
143-k7-corner	0.68E+8	0.34E+8	2.65%	1.61E+8	1370	-86.56%	-58.24	0.0000
153-k22-clustr	0.59E+8	0.27E+8	1.30%	1.64E+8	1794	-89.47%	-84.49	0.0000
162-k11-center	0.18E+8	0.10E+8	0.85%	1.62E+8	1396	-79.89%	-31.94	0.0000
256-k16-clustr	0.37E+8	0.17E+8	1.90%	2.00E+8	1647	-84.70%	-52.93	0.0000
261-k13-corner	0.41E+8	0.23E+8	2.65%	1.78E+8	1536	-93.84%	-120.83	0.0000
270-k35-center	1.12E+8	0.39E+8	3.75%	2.07E+8	1432	-89.68%	-111.92	0.0000
331-k15-corner	1.47E+8	0.51E+8	3.70%	3.15E+8	1524	-80.60%	-53.66	0.0000
344-k43-center	2.33E+8	0.87E+8	5.25%	2.73E+8	1495	-89.55%	-102.02	0.0000
351-k40-clustr	3.14E+8	0.68E+8	5.45%	2.84E+8	1651	-89.20%	-169.25	0.0000
Average**	1.14E+8	1.05E+8	3.06%	2.16E+8	1538	-88.62%	-113.61	0.0000
				Pre-Selected tALNS, ω=0.15****				
143-k7-corner	0.44E+8	0.28E+8	2.60%	1.40E+8	1551	-91.19%	-51.98	0.0000
153-k22-clustr	0.71E+8	0.16E+8	1.20%	1.84E+8	1797	-87.48%	-97.32	0.0000
162-k11-center	0.07E+8	0.06E+8	1.00%	1.41E+8	1512	-92.31%	-41.05	0.0000
256-k16-clustr	0.22E+8	0.09E+8	1.00%	1.60E+8	1552	-90.83%	-79.05	0.0000
261-k13-corner	0.24E+8	0.20E+8	2.10%	1.58E+8	1757	-96.50%	-104.27	0.0000
270-k35-center	1.26E+8	0.26E+8	3.80%	1.84E+8	1700	-88.39%	-116.60	0.0000
331-k15-corner	0.54E+8	0.42E+8	3.00%	1.66E+8	1690	-92.86%	-52.47	0.0000
344-k43-center	1.68E+8	0.34E+8	4.80%	1.98E+8	1657	-92.45%	-193.58	0.0000
351-k40-clustr	1.90E+8	0.48E+8	5.00%	2.04E+8	1701	-93.48%	-177.11	0.0000
Average*	0.78E+8	0.69E+8	2.72%	1.71E+8	1657	-92.20%	-127.22	0.0000
				LLNS				
143-k7-corner	0.29E+8	0.16E+8	1.40%	1.63E+8	1442	-94.26%	-93.91	0.0000
153-k22-clustr	0.34E+8	0.19E+8	1.10%	2.03E+8	1828	-93.97%	-88.04	0.0000
162-k11-center	0.01E+8	0.00E+8	0.50%	1.67E+8	974	-99.32%	-625.95	0.0000
256-k16-clustr	0.11E+8	0.05E+8	0.90%	1.82E+8	1674	-95.57%	-133.44	0.0000
261-k13-corner	0.15E+8	0.08E+8	2.00%	1.83E+8	1878	-97.71%	-254.27	0.0000
270-k35-center	1.58E+8	0.18E+8	2.80%	2.20E+8	1859	-85.44%	-166.44	0.0000
331-k15-corner	0.15E+8	0.13E+8	1.90%	1.94E+8	1926	-97.98%	-177.80	0.0000
344-k43-center	2.01E+8	0.29E+8	4.00%	2.30E+8	1877	-90.97%	-223.58	0.0000
351-k40-clustr	2.95E+8	0.80E+8	5.30%	2.54E+8	1843	-89.86%	-103.80	0.0000
Average*	0.84E+8	1.05E+8	2.21%	2.00E+8	1700	-91.62%	-83.13	0.0000

*10 runs per Instance, **20 runs per Instance, ***two sided p-Value
****Pre-Selected (PS) refers to a reduced operator portfolio of *greedy_*, *nearby_*, *locost_* and *random_repair*.
ω: Time Sensitivity Factor for tALNS. Higher value means lower rewards for slow operators.
n: instance size in number of customer nodes, **k**: min. amount of vehicles needed, **corner/clustr/center**: instance layout

We can deduct the following insights from the algorithm comparison results:

- Similar to the results found by Turkeš et al. (2021), our ALNS barely outperforms random operator selection (¬)ALNS by 2.85%, with an insignificant p-value of 0.786.
- The tALNS significantly outperforms ALNS: by 62.49% (p=0.000) for the complete operator portfolio, by 31.48% (p=0.000) in the pre-selected operator portfolio. This shows that the time-sensitive ALNS is a strong extension to classic ALNS, especially when operator portfolios are large and include inefficient operators. Section 4.6.2 goes into depth on the influence of the operator portfolio and the importance of including a tuned sensitivity factor ω.
- In this study, we are able to significantly outperform the ALNS with a RL-trained repair operator selection policy. We improve our objective value by 91.85%, reducing runtimes to less than 10% of ALNS. While early versions of our LLNS struggle to reach the results of ALNS with a pre-selected operator portfolio, the addition of decaying hyper-parameters and additional (yet still domain-agnostic) features allows the training to converge towards a LNS-style policy featuring mainly one very efficient operator. It is unclear whether the conditions of our problem setting and operator portfolio or insufficient training cause this result. Nevertheless, we can prove that in our setting, LLNS can replace the time-intensive and error-prone operator preselection by hand.

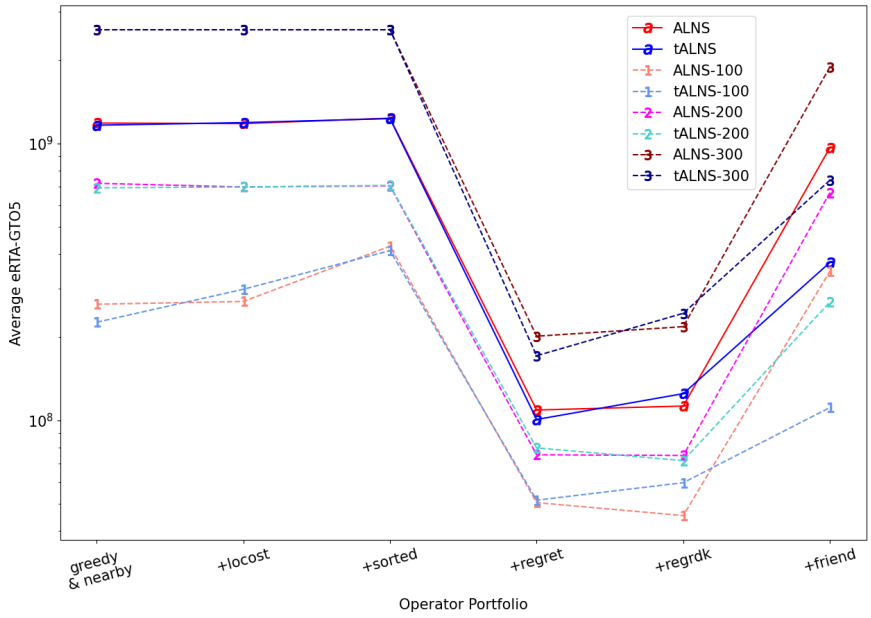

Figure 8: Performance of ALNS and ALNS with growing Operator Portfolios. Time-Sensitivity Factor ω is fixed at 1. 5 runs per algorithm, portfolio and instance.

4.6.2 Time-Sensitive ALNS: Operator Portfolio Comparison

This section compares under which boundary conditions tALNS outperform ALNS given a fixed Time-Sensitivity Factor $\omega = 1$, which is equal to not utilizing a sensitivity factor. The Literature Review Table 1 shows that out of nine sampled studies, only Gullhav et al. (2017) implement a tuneable factor to control the impact of runtime. Figure 8 shows the average eRTA-GTO5 reached by the two algorithms when provided with growing operator portfolios. We order the operators by their average iteration runtime and add a new slowest operator for each new portfolio. The figure shows average objective values over all instance centuries and an additional graph for each evaluation instance size century.

Multiple insights can be gained from this analysis: Firstly, the faster operators (up to *sorted_ repair*) are not strong enough to solve the 300-century instances to GTO10 within 2000 iterations. As no expected RTA can be calculated, we punish these runs with the slowest run from any portfolio that does reach GTO10. The exact punished and unpunished values, as well as a significance test, can be found in Appendix Table A. Secondly, tALNS strongly outperforms ALNS once the very slow *friend_repair* operator is added. This is the optimal scenario for tALNS: the presence of an operator that finds a high amount of improvements per iteration - leading to high ALNS operator weight - but performs poorly in improvements per runtime. Lastly, we notice that tALNS underperforms once the two *regret*-operators are added. This happens when the most effective operators (*regret_* and *regrdk_repair*) are also the comparatively slowest operators. As shown in Section 4.6.1, a LNS with only *regret_repair* outperforms all other algorithms, but tALNS still lowers the rewards of *regret_repair* as it is slower than the average. It therefore takes longer for the weight of the most effective operator to

Table 8: Results of Tuning Time-Sensitivity Factor ω for Pre-Selected Operator Portfolio (*greedy_, nearby_, locost_* and *regret_repair*)

ω:	0.00	0.10	0.15	0.25	0.35	0.50	0.75	1.00	1.20
	1.14E+8	0.94E+8	0.78E+8	0.88E+8	0.92E+8	0.84E+8	0.89E+8	0.92E+8	1.31E+8

ω: Time Sensitivity Factor for tALNS. At minimum 45 runs per parameter value.

Table 9: Results of Testing tALNS and LLNS in Large Instances

Instance**	Avg. eRTA-GTO5	Std. Dev. eRTA-GTO5	Avg. Final GTO	Avg. Final RTA	Avg. Iter. of Best Solution	% Change to ALNS	T-Value	p-Value****
				ALNS				
n701-k44_corner	25.97E+8	5.18E+8	6.4%	18.48E+8	1772			
n716-k35_clustr	27.31E+8	13.72E+8	6.4%	17.18E+8	1688			
n733-k159_center	-*	-*	11.4%	29.07E+8	1921			
Average* **	26.64E+8	9.81E+8	8.13%	21.57E+8	1794			
				tALNS, ω=1.2				
n701-k44_corner	6.19E+8	1.20E+8	6.2%	8.37E+8	1739	-76.15%	-36.80	0.0000
n716-k35_clustr	12.09E+8	3.63E+8	6.8%	7.99E+8	1825	-55.73%	-9.38	0.0007
n733-k159_center	-*	-*	11.4%	15.50E+8	1707	-	-	-
Average* **	9.14E+8	4.02E+8	8.13%	10.62E+8	1757	-65.68%	-16.87	0.0000
				LLNS				
n701-k44_corner	3.23E+8	0.56E+8	5.60%	2.68E+8	1918	-87.56%	-90.76	0.0000
n716-k35_clustr	3.32E+8	0.74E+8	5.20%	2.92E+8	1863	-87.86%	-72.39	0.0000
n733-k159_center	-*	-*	12.00%	10.94E+8	1709	-	-	-
Average* **	3.27E+8	0.62E+8	7.60%	5.52E+8	1830	-87.71%	-145.73	0.0000

*GTO10 was not reached, **5 runs per Instance and Algorithm, ***Runs which reached GTO10, ****two sided p-Value
n: instance size in number of customer nodes, **k**: min. amount of vehicles needed, **corner/clustr/center**: instance layout

outgrow the faster operators, leading to overall longer runtimes. For this analysis it is very important to stress that the time-sensitivity meta-parameter ω is not tuned for each portfolio. If we re-tune for each portfolio, we expect ω to be < 1 for the portfolios where the slowest operator is the most efficient. In fact, a sensitivity value of $\omega = 0$ would behave exactly like ALNS.

While Gullhav et al. (2017) tune their ω only to as low as 0.5, we test ω-values from 0 to 1.2 for the PS operator portfolio (including *greedy_, nearby_, locost_* and *regret_repair*). Table 8 shows the tuning results, and Figure 7 shows that a sensitivity factor of $\omega = 0.15$ not only outperforms ALNS, but matches the performance of LLNS. The analysis in this section shows the strengths of tALNS when utilizing a tuned time-sensitivity factor.

4.6.3 Generalization

Table 9 shows the results of testing whether or not tALNS and LLNS can generalize to larger instance sizes. We can show that for our implementation, the operator preferred by LLNS still outperforms ALNS. tALNS also still performs much better than raw ALNS. Note that none of the algorithms is able to find solutions with GTO10 for instance 733, which features a very large amount of routes (at least 159 vehicles needed to serve all customers). The algorithms reach similar solution qualities in all instances, with LLNS doing so the fastest. To further evaluate the generalization capabilities of the algorithms, tests in different problem domains are needed. This is out of the scope of this study and subject of further research.

5 Conclusion

This work tests two enhancements to the Adaptive Large Neighborhood Search: Firstly, we show that adding time-sensitivity to the weight adjustment scheme can significantly speed up the ALNS, especially when many complex operators are present. We furthermore prove that including a tuned sensitivity factor, which is currently not common in the existing literature, is crucial for tALNS to perform well even in edge cases where the most efficient operators are slower than average. By outperforming classic ALNS even with small hand-picked operator portfolios by more than 30%, we show that the inclusion of tALNS is worth the added complexity (in our case, the timing operations slow the algorithm down by roughly 2%) and the additional meta-parameter. Further work is needed to prove that tALNS performs well outside our problem domain. Additionally, testing is needed to compare how tALNS behaves with pairwise operator selection, as the degree of destruction of a given destroy operator significantly influences the runtime of the following repair operator.

The second tested enhancement replaces the adaptive layer with a reinforcement learned operator selection policy, creating the Learned Large Neighborhood Search. We are able to learn a policy that consistently chooses the most efficient repair operator and significantly outperforms ALNS even after operator pre-selection and only using domain-agnostic features. Further testing is needed to determine whether or not a more efficient policy exists that utilizes more than one operator. Our training setup leaves room for improvement. Firstly, we spend around 99% of the considerable training time solving the vehicle routing problem. We estimate that we compute more than 50 million transitions

throughout this study. By saving these transitions for an offline training dataset, training itself and especially testing training configurations can be sped up massively. Secondly, a more substantial basis for operator selection can be formed by incorporating domain-specific features. This could be especially beneficial for more complex problem domains where specialized operators are needed to handle different aspects of the problem, for example, the vehicle assignment in CVRP with heterogeneous fleets.

References

Adulyasak, Y., Cordeau, J.-F., & Jans, R. (2014). Optimization-based Adaptive Large Neighborhood Search for the Production Routing Problem. *Transportation Science*, *48*(1), 20–45. http://www.jstor.org/stable/43666994

Ahuja, R. K., Ergun, Ö., Orlin, J. B., & Punnen, A. P. (2002). A Survey of Very Large-scale Neighborhood Search Techniques. *Discrete Applied Mathematics*, *123*(1-3), 75–102. https://doi.org/10.1016/S0166-218X(01)00338-9

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A Distributional Perspective on Reinforcement Learning. http://arxiv.org/pdf/1707.06887v1

Ben-Ameur, W. (2004). Computing the Initial Temperature of Simulated Annealing. *Computational Optimization and Applications*, *29*(3), 369–385. https://doi.org/10.1023/B:COAP.0000044187.23143.bd

Bengio, Y. (2012). Practical Recommendations for Gradient-based Training of Deep Architectures. https://doi.org/10.48550/arXiv.1206.5533

Bengio, Y., Lodi, A., & Prouvost, A. (2021). Machine Learning for Combinatorial Optimization: A Methodological Tour D'horizon. *European Journal of Operational Research*, *290*(2), 405–421. https://doi.org/10.1016/j.ejor.2020.07.063

Bongiovanni, C., Kaspi, M., Cordeau, J.-F., & Geroliminis, N. (2022). A Machine Learning-driven Two-phase Metaheuristic for Autonomous Ridesharing Operations. *Transportation Research Part E: Logistics and Transportation Review*, *165*. https://doi.org/10.1016/j.tre.2022.102835

Boualamia, H., Metrane, A., Hafidi, I., & Mellouli, O. (2023). A New Adaptation Mechanism of the ALNS Algorithm Using Reinforcement Learning. In N. Aboutabit, M. Lazaar, & I. Hafidi (Eds.), *Advances in Machine Intelligence and Computer Science Applications* (pp. 3–14, Vol. 656). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-29313-9{\textunderscore}1

Canca, D., De-Los-Santos, A., Laporte, G., & Mesa, J. A. (2018). The Railway Network Design, Line Planning and Capacity Problem: an Adaptive Large Neighborhood Search Metaheuristic. In J. Żak, Y. Hadas, & R. Rossi (Eds.), *Advanced Concepts, Methodologies and Technologies for Transportation and Logistics* (pp. 198–219, Vol. 572). Springer International Publishing. https://doi.org/10.1007/978-3-319-57105-8{\textunderscore}9

Černý, V. (1985). Thermodynamical Approach to the Traveling Salesman Problem: an Efficient Simulation Algorithm. *Journal of Optimization Theory and Applications*, *45*(1), 41–51. https://doi.org/10.1007/BF00940812

Chen, X., & Tian, Y. (2018). Learning to Perform Local Rewriting for Combinatorial Optimization. https://doi.org/10.48550/arXiv.1810.00337

Da Costa, P., Rhuggenaath, J., Zhang, Y., Akcay, A., & Kaymak, U. (2021). Learning 2-Opt Heuristics for Routing Problems via Deep Reinforcement Learning. *SN Computer Science*, *2*(5). https://doi.org/10.1007/s42979-021-00779-2

Dantzig, G. B., & Ramser, J. H. (1959). The Truck Dispatching Problem. *Management Science*, *6*(1), 80–91. https://doi.org/10.1287/mnsc.6.1.80

Developers, T. (2023). Tensorflow. https://doi.org/10.5281/zenodo.4724125

Ferrari, S., & Stengel, R. F. (2005). Smooth Function Approximation Using Neural Networks. *IEEE Transactions on Neural Networks*, *16*(1), 24–38. https://doi.org/10.1109/TNN.2004.836233

Gao, L., Chen, M., Chen, Q., Luo, G., Zhu, N., & Liu, Z. (2020). Learn to Design the Heuristics for Vehicle Routing Problem. https://doi.org/10.48550/arXiv.2002.08539

Gendreau, M., & Potvin, J.-Y. (2010). *Handbook of Metaheuristics* (Vol. 146). Springer US. https://doi.org/10.1007/978-1-4419-1665-5

Godinho, M. T., Gouveia, L., & Magnanti, T. L. (2008). Combined Route Capacity and Route Length Models for Unit Demand Vehicle Routing Problems. *Discrete Optimization*, *5*(2), 350–372. https://doi.org/10.1016/j.disopt.2007.05.001

Gullhav, A. N., Cordeau, J.-F., Hvattum, L. M., & Nygreen, B. (2017). Adaptive Large Neighborhood Search Heuristics for Multi-tier Service Deployment Problems in Clouds. *European Journal of Operational Research*, *259*(3), 829–846. https://doi.org/10.1016/j.ejor.2016.11.003

Hochba, D. S. (1997). Approximation Algorithms for Np-hard Problems. *ACM SIGACT News*, *28*(2), 40–52. https://doi.org/10.1145/261342.571216

Hottung, A., & Tierney, K. (2022). Neural Large Neighborhood Search for Routing Problems. *Artificial Intelligence*, *313*, 103786. https://doi.org/10.1016/j.artint.2022.103786

Joe, W., & Lau, H. C. (2020). Deep Reinforcement Learning Approach to Solve Dynamic Vehicle Routing Problem with Stochastic Customers. *Proceedings of the International Conference on Automated Planning and Scheduling*, *30*, 394–402.

Johnn, S.-N., Darvariu, V.-A., Handl, J., & Kalcsics, J. (2023). Graph Reinforcement Learning for Operator Selection in the ALNS Metaheuristic. https://doi.org/10.48550/arXiv.2302.14678

Karimi-Mamaghan, M., Mohammadi, M., Meyer, P., Karimi-Mamaghan, A. M., & Talbi, E.-G. (2022). Machine Learning at the Service of Meta-heuristics for Solving Combinatorial Optimization Problems: A State-of-the-art. *European Journal of Operational Research*, *296*(2), 393–422. https://doi.org/10.1016/j.ejor.2021.04.032

Kerschke, P., Hoos, H. H., Neumann, F., & Trautmann, H. (2019). Automated Algorithm Selection: Survey and Perspectives. *Evolutionary computation*, *27*(1), 3–45. https://doi.org/10.1162/evco{\textunderscore}a{\textunderscore}00242

Kiefer, A., Hartl, R. F., & Schnell, A. (2017). Adaptive Large Neighborhood Search for the Curriculum-based Course Timetabling Problem. *Annals of Operations Research*, *252*(2), 255–282. https://doi.org/10.1007/s10479-016-2151-2

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science (New York, N.Y.)*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Kool, W., van Hoof, H., & Welling, M. (2018). Attention, Learn to Solve Routing Problems! In *International Conference on Learning Representations*. ICLR. https://doi.org/10.48550/arXiv.1803.08475

Kucukoglu, I., Dewil, R., & Cattrysse, D. (2021). The Electric Vehicle Routing Problem and Its Variations: A Literature Review. *Computers & Industrial Engineering*, *161*, 107650. https://doi.org/10.1016/j.cie.2021.107650

Laborie, P., & Godard, D. (2007). Self-adapting Large Neighborhood Search: Application to Single-mode Scheduling Problems. In *Proceedings MISTA-07, Paris, Vol. 8* (Vol. 8). Multidisciplinary International Conference on Scheduling.

Lima, I., Oliveira, D., & Queiroga, E. (2023). The Capacitated Vehicle Routing Problem Library (E. Queiroga, Ed.). Retrieved November 24, 2023, from http://vrp.galgos.inf.puc-rio.br/index.php/en/plotted-instances

Lu, H., Zhang, X., & Yang, S. (2020). A Learning-based Iterative Method for Solving Vehicle Routing Problems. In *International Conference on Learning Representation 2020*. Retrieved April 12, 2023, from https://openreview.net/forum?id=BJe1334YDH

Nazari, M., Oroojlooy, A., Snyder, L. V., & Takáč, M. (2018). Reinforcement Learning for Solving the Vehicle Routing Problem. https://doi.org/10.48550/arXiv.1802.04240

Oliva, D., Houssein, E. H., & Hinojosa, S. (2021). *Metaheuristics in Machine Learning: Theory and Applications*. Springer.

Pisinger, D., & Ropke, S. (2007). A General Heuristic for Vehicle Routing Problems. *Computers & Operations Research*, *34*(8), 2403–2435. https://doi.org/10.1016/j.cor.2005.09.012

Reijnen, R., Zhang, Y., Lau, H. C., & Bukhsh, Z. (2022). Operator Selection in Adaptive Large Neighborhood Search Using Deep Reinforcement Learning. https://doi.org/10.48550/arXiv.2211.00759

Ropke, S., & Pisinger, D. (2006). An Adaptive Large Neighborhood Search Heuristic for the Pickup and Delivery Problem with Time Windows. *Transportation Science*, *40*(4), 455–472. https://doi.org/10.1287/trsc.1050.0135

Santini, A., Ropke, S., & Hvattum, L. M. (2018). A Comparison of Acceptance Criteria for the Adaptive Large Neighbourhood Search Metaheuristic. *Journal of Heuristics*, *24*(5), 783–815. https://doi.org/10.1007/s10732-018-9377-x

Shaw, P. (1998). Using Constraint Programming and Local Search Methods to Solve Vehicle Routing Problems. In G. Goos, J. Hartmanis, J. van Leeuwen, M. Maher, & J.-F. Puget (Eds.), *Principles and Practice of Constraint Programming — CP98* (pp. 417–431, Vol. 1520). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-49481-2{\textunderscore}30

Song, J., Lanka, R., Yue, Y., & Dilkina, B. (2020). A General Large Neighborhood Search Framework for Solving Integer Linear Programs. https://doi.org/10.48550/arXiv.2004.00422

Sonnerat, N., Wang, P., Ktena, I., Bartunov, S., & Nair, V. (2021). Learning a Large Neighborhood Search Algorithm for Mixed Integer Programs. https://doi.org/10.48550/arXiv.2107.10201

Stathakis, D. (2009). How Many Hidden Layers and Nodes? *International Journal of Remote Sensing*, *30*(8), 2133–2147. https://doi.org/10.1080/01431160802549278

Thomas, C., & Schaus, P. (2018). Revisiting the Self-adaptive Large Neighborhood Search. In W.-J. van Hoeve (Ed.), *Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (pp. 557–566, Vol. 10848). Springer International Publishing. https://doi.org/10.1007/978-3-319-93031-2{\textunderscore}40

Turkeš, R., Sörensen, K., & Hvattum, L. M. (2021). Meta-analysis of Metaheuristics: Quantifying the Effect of Adaptiveness in Adaptive Large Neighborhood Search. *European Journal of Operational Research*, *292*(2), 423–442. https://doi.org/10.1016/j.ejor.2020.10.045

Uchoa, E., Pecin, D., Pessoa, A., Poggi, M., Vidal, T., & Subramanian, A. (2017). New Benchmark Instances for the Capacitated Vehicle Routing Problem. *European Journal of Operational Research*, *257*(3), 845–858. https://doi.org/10.1016/j.ejor.2016.08.012

Wang, P., Reinelt, G., & Tan, Y. (2012). Self-adaptive Large Neighborhood Search Algorithm for Parallel Machine Scheduling Problems. *Journal of Systems Engineering and Electronics*, *23*(2), 208–215. https://doi.org/10.1109/JSEE.2012.00027

Windras Mara, S. T., Norcahyo, R., Jodiawan, P., Lusiantoro, L., & Rifai, A. P. (2022). A Survey of Adaptive Large Neighborhood Search Algorithms and Applications. *Computers & Operations Research*, *146*, 105903. https://doi.org/10.1016/j.cor.2022.105903

Wu, Y., Yang, W., He, G., & Zhao, S. (2017). An Improved Adaptive Large Neighborhood Search Algorithm for the Heterogeneous Fixed Fleet Vehicle Routing Problem. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 657–663. https://doi.org/10.1109/ICSESS.2017.8343000

Wu, Y., Song, W., Cao, Z., & Zhang, J. (2021). Learning Large Neighborhood Search Policy for Integer Programming. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, & J. Wortman Vaughan (Eds.), *Advances in Neural Information Processing Systems* (pp. 30075–30087, Vol. 34). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/fc9e62695def29ccdb9eb3fed5b4c8c8-Paper.pdf

Wu, Y., Song, W., Cao, Z., Zhang, J., & Lim, A. (2022). Learning Improvement Heuristics for Solving Routing Problems. *IEEE transactions on neural networks and learning systems*, *33*(9), 5057–5069. https://doi.org/10.1109/TNNLS.2021.3068828

Zhang, H., & Yang, W. (2021). An Enhanced Adaptive Large Neighborhood Search Algorithm for the Capacitated Vehicle Routing Problem. *2021 13th International Conference on Machine Learning and Computing*, 79–85. https://doi.org/10.1145/3457682.3457694

Zhou, J., Wu, Y., Cao, Z., Song, W., Zhang, J., & Chen, Z. (2023). Learning Large Neighborhood Search for Vehicle Routing in Airport Ground Handling. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 9769–9782. https://doi.org/10.1109/TKDE.2023.3249799

List of Abbreviations

(¬)ALNS non-adaptive Large Neighborhood Search 2, 12, 14, 16, 17, 19–21

ALNS Adaptive Large Neighborhood Search 1–3, 6–11, 13–15, 17–22

ANN Artificial Neural Network 1, 2, 14

CO Combinatorial Optimization 1, 2, 6

CVRP Capacitated Vehicle Routing Problem 1, 2, 6–8, 23

DQL Deep-Q Learning 2, 7

DQN Deep-Q Network 13

DRL Deep Reinforcement Learning 6

eRTA expected runtime approximation 15, 17

eRTA-GTO5 expected runtime approximation value at 5% gap to optimum 13–19, 21

GTO gap to optimum 5, 13–17, 19, 21, 22

GTO5 5% gap to optimum 13, 14

LLNS Learned Large Neighborhood Search 2, 7, 8, 11, 13, 14, 17–22

LNS Large Neighborhood Search 2, 6, 8, 16, 18, 21

MDP Marcov Decision Process 6

ML Machine Learning 2, 6, 8

ns nano-second 15

OR Operations Research 1, 2, 6

PS pre-selected 18, 22

RL Reinforcement Learning 17, 21

RTA runtime approximation 11–15, 17–19, 21

SA Simulated Annealing 6, 8, 11, 13, 14, 17

SA-LNS Self-Adapting Large Neighborhood Search 3, 6

tALNS runtime-sensitive Adaptive Large Neighborhood Search 1, 2, 7, 8, 14, 17–22

TSP Traveling Salesman Problem 6

VRP Vehicle Routing Problem 2, 6, 8

Notation

α **update rate** for operator weights 4, 17

b **batch** of iterations 4, 8, 11

β **length of batch** in iterations 4, 8, 11, 17

c **cost** of solution $S \rightarrow c_S$ 10, 13

ct **cumulative runtimes** of operator o in batch $b \rightarrow ct_{o,b}$ 8

dod **degree of destruction** of destroy operator o, the percentage of nodes to remove out of the solution $\rightarrow dod_o$ 14, 17

ϵ **epsilon-greedy** factor deciding how often a policy should select a random action during training 14, 17, 18

gf **gap factor** for scaling the Gap to Optimum in the LLNS reward function 13, 15, 17

i **iteration** 8

ω **time sensitivity factor** used to scale the impact of runtime on tALNS weights 4, 7, 11, 17, 19–22

Ω^- **set** of destroy operators 7, 8

Ω^+ **set** of repair operators 7, 8, 11

p **probability** of operator being chosen in batch $b \rightarrow p_{o,b}$ 10, 11

ϕ **reward** for the acceptance check with outcome $\rho \rightarrow \phi_\rho$ 4, 8, 11, 13, 17

Φ **score** of operator o in iteration $i \rightarrow \Phi_{o,i}$ 8, 11

R **reward** gained for a given transition in Reinforcement Learning 13

ρ **result** of the acceptance check in iteration $i \rightarrow \rho_i$ 8, 11, 13, 17

S **active solution** 8, 10, 13

S^* **best solution** 8, 13

S^a **accepted solution** 8, 10, 13

sa_cr simulated annealing **cooling rate** to lower temperature 10, 11

sat simulated annealing **temperature** at iteration $i \rightarrow sat_i$ 8, 10

sa_ti simulated annealing **target iteration** for calculating cooling rate 10, 11

sa_tp simulated annealing **target percentage** for calculating cooling rate 10, 11

$S^{opt.}$ **optimal solution** 13

Θ **number of appearances** of destroy or repair operator o in batch $b \rightarrow \Theta_{o,b}^{-/+}$ 4, 8, 11

t **runtime** of destroy or repair operator o in iteration $i \rightarrow t_{o,i}^{-/+}$ 4, 8, 11

tf **time factor** for tALNS 4, 11

w **weight** of operator o in batch $b \rightarrow w_{o,b}$ 4, 8, 11

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Die Rolle von Corporate Digital Responsibility im digitalen Controlling

David Schopp

Heinrich Heine University Düsseldorf

Abstract

Digitalization is fundamentally changing controlling and, at the same time, increasing the demands for responsible handling of data, technologies, and automated decision-making systems. Against this backdrop, this thesis examines the extent to which CDR influences the design and implementation of digital controlling processes. The aim of the thesis is to systematically present the influence of ethical, data protection and technological aspects on digital controlling and to identify practical starting points for the integration of CDR. Methodologically, the thesis is based on a literature analysis of relevant scientific articles as well as regulatory and practical sources. Building on the domains of digital controlling according to Keimer and Egle, the effects of CDR on the areas of data, technologies, processes, methods and competencies are analyzed. The focus is particularly on data governance, data protection, artificial intelligence and big data, as well as the changing role of the controller. The results show that CDR as a framework concept permeates all domains of digital controlling and significantly expands the requirements for controlling systems and controller competencies. The responsible use of digital technologies requires transparent data structures, ethically reflective analysis methods and the strategic anchoring of CDR in corporate management. This work thus contributes to closing an existing research gap and illustrates that CDR is not only a normative obligation but can also create sustainable added value for companies and stakeholders.

Zusammenfassung

Die Digitalisierung verändert das Controlling grundlegend und erhöht zugleich die Anforderungen an einen verantwortungsvollen Umgang mit Daten, Technologien und automatisierten Entscheidungssystemen. Vor diesem Hintergrund untersucht diese Arbeit, inwiefern CDR die Gestaltung und Umsetzung digitaler Controllingprozesse beeinflusst. Ziel der Arbeit ist es, den Einfluss ethischer, datenschutzrechtlicher und technologischer Aspekte auf das digitale Controlling systematisch darzustellen und praxisorientierte Ansatzpunkte für die Integration von CDR aufzuzeigen. Methodisch basiert die Arbeit auf einer Literaturanalyse einschlägiger wissenschaftlicher Beiträge sowie regulatorischer und praxisnaher Quellen. Aufbauend auf den Domänen des digitalen Controllings nach Keimer und Egle werden die Auswirkungen von CDR auf die Bereiche Daten, Technologien, Prozesse, Methoden und Kompetenzen analysiert. Dabei stehen insbesondere Data Governance, Datenschutz, Künstliche Intelligenz und Big Data sowie der Wandel der Controllerrolle im Fokus. Die Ergebnisse zeigen, dass CDR als Rahmenkonzept alle Domänen des digitalen Controllings durchdringt und die Anforderungen an Controllingsysteme sowie Controllerkompetenzen deutlich erweitert. Ein verantwortungsvoller Einsatz digitaler Technologien erfordert transparente Datenstrukturen, ethisch reflektierte Analyseverfahren und eine strategische Verankerung von CDR in der Unternehmensführung. Diese Arbeit leistet damit einen Beitrag zur Schließung einer bestehenden Forschungslücke und verdeutlicht, dass CDR nicht nur eine normative Verpflichtung darstellt, sondern auch einen nachhaltigen Mehrwert für Unternehmen und Stakeholder schaffen kann.

Keywords: corporate digital responsibility; digitales controlling; künstliche intelligenz; datenschutz

1 Einleitung

Die digitale Transformation verändert die Welt schneller und tiefgreifender als es in den vergangenen Jahrzehnten vorstellbar war. Unternehmen stehen vor der Herausforderung, effizienter zu arbeiten und dabei auch ethische sowie gesellschaftliche Verantwortung zu übernehmen. Denn Technologien wie Künstliche Intelligenz (KI), Big Data und automatisierte Systeme ermöglichen neue Ansätze in der Unternehmenssteuerung, erfordern jedoch ein Umdenken in Bezug auf den Umgang mit Daten, Transparenz und Nachhaltigkeit. Besonders im Bereich des Controllings, der als zentrale Informations- und Steuerungseinheit in Unternehmen dient, wird die Frage nach einer verantwortungsvollen Nutzung digitaler Technologien immer wichtiger. Hier setzt das Konzept der Corporate Digital Responsibility (CDR) an, das die gesellschaftliche Verantwortung von Unternehmen im digitalen Kontext definiert und erweitert.

Ein wesentlicher Aspekt dieser Verantwortung liegt im digitalen Controlling, das immer mehr auf datengetriebene Technologien zurückgreift. Dabei geht es nicht nur darum,

DOI: https://doi.org/10.5282/jums/v11i1pp27-42

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

Unternehmenszahlen präzise zu erfassen und zu analysieren, sondern auch wie diese Daten gesammelt, gespeichert und verwendet werden. Big Data beschreibt in diesem Zusammenhang die Fähigkeit, riesige Datenmengen in Echtzeit zu sammeln, zu speichern und zu analysieren vgl. (Dörr, 2020, p. 14). Dies wäre manuell kaum möglich und erfordert leistungsstarke Technologien und Algorithmen. Hier kommen datenschutzrechtliche Fragestellungen und ethische Überlegungen ins Spiel: Wie kann der Einsatz von KI und Algorithmen im Controlling so gestaltet werden, sodass Fairness, Transparenz und Datenschutz gewährleistet sind? Die Corporate Social Responsibility (CSR) hat bereits ähnliche Fragen in Bezug auf Umwelt- und Sozialverantwortung beantwortet. Nun gilt es, diese Konzepte in den digitalen Raum zu übertragen und auf die Herausforderungen der modernen Unternehmensführung anzuwenden.

Das Thema ist auch aus einer wirtschaftlichen Perspektive von großer Bedeutung. Unternehmen, die CDR ignorieren, setzen sich nicht nur rechtlichen Risiken aus, sondern können auch ihr Image und das Vertrauen sowohl bei den Verbrauchern als auch anderer Stakeholder wie Kapitalgeber, Mitarbeitende oder die Politik aufs Spiel setzen vgl. (Dörr, 2020, p. 49). Die Umsetzung und strategische Einbindung von CDR in der Unternehmensführung bieten hingegen die Chance, Wettbewerbsvorteile durch eine bessere Reputation zu erzielen und das Vertrauen zu stärken vgl. (Weißenberger & Marrocco, 2022, p. 47). Darüber hinaus trägt CDR dazu bei, Unternehmen resilienter gegen zukünftige Herausforderungen zu machen, sei es durch neue Datenschutzgesetze oder den wachsenden Druck der Gesellschaft, ethisch verantwortungsvoll zu handeln.

Die Forschung zu CDR befindet sich derzeit noch in einem frühen Stadium. Zwar gibt es viele Ansätze, die digitale Verantwortung von Unternehmen, insbesondere gegenüber Verbrauchern zu definieren, doch eine detaillierte Betrachtung über die Auswirkungen von CDR auf das digitale Controlling und die Einbindung der Stakeholder der Unternehmenssteuerung fehlt weitgehend. Gleichzeitig zeigt eine Analyse der 100 umsatzstärksten Unternehmen in Deutschland, dass 15 % sich öffentlich mit dem Thema CDR auseinandersetzen, während 91 % lediglich CSR thematisieren vgl. (Andersen, 2022, p. 9). Eine weitere Analyse von nicht-finanziellen Berichten der DAX-30-Unternehmen veranschaulicht, dass der Großteil der CDR-Aktivitäten kunden- und mitarbeiterbezogen ist, während leistungs- und umweltbezogene Themen eine geringere Rolle spielen vgl. (Merbecks, 2023). Zudem variiert die Relevanz von CDR stark zwischen den Unternehmen, was eine erhebliche Lücke in der Unternehmenskommunikation und eine mangelnde Verankerung der digitalen Verantwortung im Bewusstsein vieler Firmen verdeutlicht.

Diese Arbeit soll folglich einen Beitrag zur wissenschaftlichen Auseinandersetzung mit CDR leisten, indem sie untersucht, wie CDR das digitale Controlling beeinflusst und welche Maßnahmen Unternehmen ergreifen können, um diese Verantwortung in ihre Prozesse zu integrieren. Die zentrale Forschungsfrage lautet daraus resultierend: „Inwiefern beeinflusst Corporate Digital Responsibility die Gestaltung und Umsetzung digitaler Controllingprozesse in Unternehmen und welche Rolle spielen dabei ethische, datenschutzrechtliche und technologische Aspekte? “. Diese Forschungsfrage zielt darauf ab, den Einfluss von CDR auf die täglichen Prozesse des digitalen Controllings zu untersuchen und dabei sowohl die Chancen als auch die Risiken zu beleuchten. Es soll klar herausgearbeitet werden, wie Unternehmen diese Verantwortung konkret in ihre Controllingprozesse integrieren können.

Es wird in der Arbeit gezeigt, dass digitale Verantwortung nicht nur eine strategische Managementaufgabe ist, sondern auch das Controlling auf seinen verschiedenen Ebenen betrifft. Die digitale Transformation des Controllings, verbunden mit der Verschiebung von Kompetenzanforderungen und Verantwortungsbereichen, macht ein CDR-orientiertes Denken und Handeln im Controlling zunehmend wichtiger. Darüber hinaus wird verdeutlicht, dass ein verantwortungsvoller und nachhaltiger Umgang mit Technologien in der Controllerarbeit, der internen sowie externen Unternehmenskommunikation und im Umgang mit den Stakeholdern unverzichtbar ist. Verantwortung in rechtlichen Verpflichtungen wie Datenschutz und ein klar definiertes ethisches Leitbild müssen integraler Bestandteil der Controlling-Arbeit sein, um regulatorische und gesellschaftliche Anforderungen zu erfüllen. Die Arbeit zeigt außerdem, dass spezifische Herausforderungen bei der Implementierung von CDR bestehen. Um diese zu bewältigen, sind praxisorientierte Ansätze erforderlich, die eine Umsetzung von CDR auf operativer und strategischer Ebene ermöglichen und so zu einem nachhaltigen Unternehmenserfolg beitragen können.

Die Arbeit gliedert sich in mehrere Kapitel: Nach dieser Einleitung werden im zweiten Kapitel die theoretischen Grundlagen von CDR und des digitalen Controllings erläutert. Dabei soll neben begrifflichen Definitionen vor allem herausgearbeitet werden, wie CDR in den Kontext von CSR gesetzt werden kann und wie die Digitalisierung die Anforderungen an das klassische Controlling verändert. Das dritte Kapitel widmet sich den spezifischen Auswirkungen von CDR auf die einzelnen von (Keimer & Egle, 2020, p. 8) definierten Domänen des digitalen Controllings wie Daten, Technologien, Prozesse, Methoden und Kompetenzen. In diesem Zusammenhang werden die Herausforderungen von CDR im Umgang mit KI und Big Data untersucht, wobei auch der Aspekt des Datenschutzes näher betrachtet wird. Im vierten Kapitel werden anhand des CDR-Kompasses nach (Weißenberger & Marrocco, 2022, p. 45) und des 6-Schritte-Modells nach (Dörr, 2020, p. 52) zwei konkrete Strategien und Maßnahmen zur Implementierung von CDR vorgestellt und miteinander verknüpft. Dabei werden mögliche Herausforderungen, die bei der Umsetzung entstehen können, aufgezeigt und Lösungsansätze vorgestellt. Abschließend werden die Ergebnisse zusammengefasst und ein Ausblick auf zukünftige Forschungs- und Praxisfelder gegeben.

2 Grundlagen der CDR und des digitalen Controllings

2.1 CDR – Ursprung, Definition und Konzepte

Seit den 1950er Jahren hat sich die Corporate Social Responsibility (CSR) als zentrales Konzept in der Unternehmensführung etabliert. Howard R. Bowen legt als Pionier mit seinem Werk „Social Responsibilities of the Businessman“ (1953) den Grundstein für das moderne Verständnis von CSR vgl. (Agudelo et al., 2019, p. 3f. Carroll, 1999, p. 270). (Bowen, 1953, p. 6) definiert diese soziale Verantwortung als „*the obligations of businessmen to pursue those policies, to make those decisions, or to follow those lines of action which are desirable in terms of the objectives and values of our society.*”. Sein Ansatz betont, dass Unternehmen nicht nur den Gewinn maximieren sollen, sondern auch soziale Verantwortung gegenüber aller Stakeholder und damit der Gesellschaft tragen und dies bei unternehmerischen Entscheidungen berücksichtigen müssen vgl. (Bowen, 1953, p. 3). Diese Perspektive war revolutionär, da sie das traditionelle, rein ökonomische Unternehmensmodell in Frage stellte und eine breitere Verantwortung einforderte vgl. (Carroll, 1999, p. 270).

Nach der Einführung des Konzepts über CSR durch Bowen hat sich dieses in den folgenden Jahrzehnten kontinuierlich weiterentwickelt, wobei sowohl akademische Beiträge als auch öffentliche und internationale Ereignisse maßgeblich dazu beigetragen haben, die gesellschaftlichen Erwartungen an unternehmerisches Verhalten zu prägen vgl. (Agudelo et al., 2019, p. 19). Ein zentrales Modell, das im Laufe dieser Entwicklung an Bedeutung gewonnen hat, ist die von (Carroll, 1991) entwickelte CSR-Pyramide (Abb. 1), die CSR als umfassendes Konzept darstellt und Unternehmen dazu anhält, wirtschaftliche, rechtliche, ethische und philanthropische Verantwortungen wahrzunehmen. Folgend werden die vier Verantwortungen kurz dargestellt.

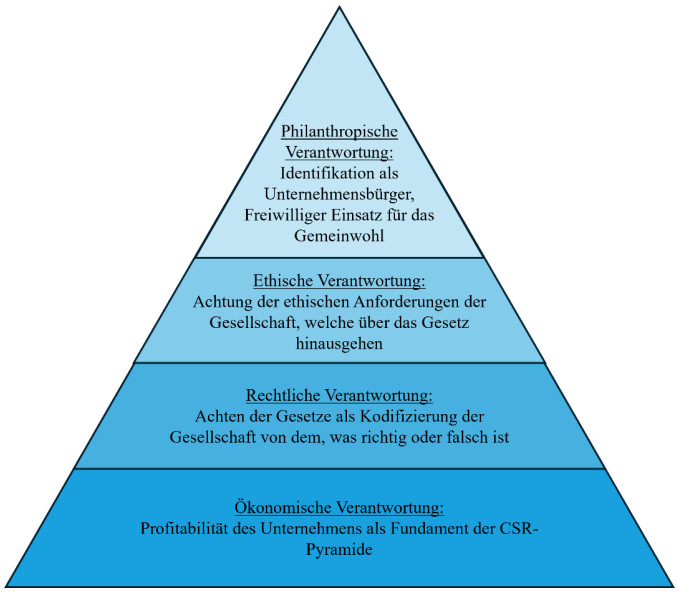

Abbildung 1: Die Pyramide der Corporate Social Responsibility
Quelle: leicht verändert entnommen aus (Carroll, 1991, p. 42)

Die ökonomische Verantwortung bildet das Fundament der Pyramide, da Unternehmen als Grundvoraussetzung profitabel sein müssen, um überhaupt existieren zu können. Ohne wirtschaftlichen Erfolg und eine gute Positionierung im Wettbewerb ist die Erfüllung der übrigen Verantwortlichkeiten nicht möglich vgl. (Carroll, 1991, p. 40f)..

Darauf aufbauend befindet sich die rechtliche Verantwortung, welche die Einhaltung von Gesetzen umfasst, die den rechtlichen Rahmen für ein verantwortungsbewusstes unternehmerisches Handeln setzen. (Carroll, 1991, p. 41) betont, dass rechtliche Verantwortungen die kodifizierte Ethik der Gesellschaft widerspiegeln, d.h. die Erwartungen der Gesellschaft an faire und gerechte Geschäftspraktiken.

Die ethische Verantwortung bildet die nächste Ebene der CSR-Pyramide. Diese geht über die gesetzlichen Anforderungen hinaus und umfasst die moralischen Erwartungen der Gesellschaft an Unternehmen. Hier geht es darum, faire, gerechte und transparente Entscheidungen zu treffen, auch wenn keine rechtlichen Verpflichtungen bestehen. Laut (Carroll, 1991, p. 41) spiegeln ethische Verantwortungen auch die sich wandelnden Werte und Normen der Gesellschaft wider, welche oftmals wiederum gesetzliche Veränderungen vorantreiben. Eine Sensibilisierung der Entscheidungsträger auf sich entwickelnde ethische Grundsätze ist somit unabdingbar.

An der Spitze der Pyramide steht die philanthropische Verantwortung, wobei es sich um freiwillige soziale Aktivitäten der Unternehmen handelt, wie z.B. Spenden oder die Unterstützung von Gemeinschaftsprojekten. Diese Handlungen sind nicht zwingend erforderlich, aber sie tragen dazu bei, das Ansehen des Unternehmens zu verbessern und eine positive Beziehung zu den Stakeholdern zu pflegen vgl. (Carroll, 1991, p. 42).

Die CSR-Pyramide geht somit über den reinen Shareholder-Value-Ansatz, also die Maximierung des Unternehmenswertes als übergeordnete Verantwortung hinaus, indem Unternehmen als aktive gesellschaftliche Akteure dargestellt werden, was an das Prinzip der Corporate Citizenship anknüpft vgl. (Schäffer & Weber, 2016b, p. 183). Corporate Citizenship fordert Unternehmen auf, über ihre ökonomischen und rechtlichen Verpflichtungen hinaus Verantwortung zu übernehmen und positive gesellschaftliche Beiträge zu leisten (Carroll, 1999, p. 289). Damit wird die Rolle von Unternehmen nicht mehr nur auf externe ökonomische Ziele reduziert, sondern in einen umfassenderen Kontext gesellschaftlicher Verantwortung eingebettet, der auch ethische und philanthropische Aspekte einschließt vgl. (Van der Merwe und Al Achkar, 2022, p. e12-2; Carroll, 1991, p. 42).

Die Bedeutung von CSR als grundlegendes Konzept verantwortungsvoller Unternehmensführung ist unbestritten. Es zielt darauf ab, soziale, ökologische und ökonomische Aspekte des unternehmerischen Handelns nachhaltig zu gestalten vgl. (Van der Merwe & Al Achkar, 2022, p. e12-2). Angesichts der digitalen Transformation und der zunehmenden Verflechtung von Technologie und Gesellschaft stellt sich die Frage, wie diese Prinzipien auf den digitalen Raum übertragen wer-

den können vgl. (Weißenberger & Marrocco, 2022, p. 44).

CDR erweitert die CSR-Prinzipien und überträgt sie damit auf die Herausforderungen und Chancen, die sich aus der Nutzung digitaler Technologien ergeben vgl. (Willrich, 2018, p. 8). Dabei liegt der Fokus auf Themen wie Datenschutz, ethischer KI-Nutzung und der Minimierung negativer gesellschaftlicher Effekte der Digitalisierung, während gleichzeitig die Potenziale digitaler Innovationen nachhaltig genutzt werden sollen vgl. (Van der Merwe und Al Achkar, 2022, p. e12-5; Pick und Knopf, 2023, p. 676f).. CDR-Maßnahmen lassen sich aus diesem Grund in die ethische Verantwortung nach (Carroll, 1991) eingliedern.

Die Etablierung einer präzisen und einheitlichen Definition von CDR erweist sich als herausfordernd, da die in der Literatur vorhandenen Ansätze häufig auf allgemeinen Formulierungen beruhen vgl. (Pick & Knopf, 2023, p. 675). Nach den Autoren beziehen sich diese Definitionen in der Regel auf nicht detailliert ausgearbeitete Normen, was eine klare Konkretisierung der Verantwortung erschwert. Ferner greifen viele Definitionen auf vage Begriffe wie Kultur zurück, die eine eindeutige und praxisorientierte Anwendung der Verantwortungskonzepte erschweren und somit die Klarheit im Verständnis von CDR begrenzen.

Das Bundesministerium für Umwelt und Verbraucherschutz (BMUV) befasste sich 2018 mit diesem Thema und leitete zusammen mit Unternehmen aus dem Technologie- sowie Handelssektor wie z.B. SAP und Zalando eine CDR-Initiative zur Etablierung eines menschen- und werteorientierten Selbstverständnisses der Unternehmensverantwortung im digitalen Raum ein[1]. Die CDR-Initiative definiert CDR *„als freiwillige unternehmerische Aktivitäten, die insbesondere im Sinne der Verbraucherinnen und Verbraucher über das gesetzlich Vorgeschriebene hinausgehen und die digitale Welt aktiv zum Vorteil der Gesellschaft mitgestalten“* (Corporate Digital Responsibility-Initiative, 2021, p. 2), welche sich in den Geschäftsberichten einiger der CDR-Initiative angehörigen Handelsunternehmen etabliert hat. Für die vorliegende Arbeit wird die Definition von CDR nach (Weißenberger & Marrocco, 2022, p. 45) zugrunde gelegt, die CDR als *„freiwillige unternehmerische Ausrichtung zur verantwortlichen Nutzung digitaler Technologien“* versteht. Diese Definition ist von Vorteil, da sie den Verantwortungsumfang nicht ausschließlich auf Verbraucher begrenzt, sondern auch andere Stakeholder des Unternehmens miteinbezieht. Darüber hinaus betont sie mit dem Terminus „unternehmerische Ausrichtung“ die Integration von CDR in der strategischen Unternehmensführung und greift im Allgemeinen Aspekte der Corporate Citizenship bei der Nutzung digitaler Technologien auf, was eine umfassendere Perspektive auf die Verantwortlichkeiten eines Unternehmens ermöglicht. Dies steht in direktem Zusammenhang mit den thematischen Schwerpunkten dieser Arbeit, insbesondere im Hinblick auf die Integration von CDR im digitalen Controlling.

[1] Vgl. https://www.bmuv.de/themen/verbraucherschutz/digitaler-verbraucherschutz/corporate-digital-responsibility, letzter Zugriff am 26.11.2024

2.2 Digitale Transformation des klassischen Controllings

Die folgenden Ausführungen zu den Kernaufgaben des klassischen Controllings stützen sich auf Darstellungen von (Binder, 2017). Das Controlling übernimmt eine zentrale Rolle in der Unternehmenssteuerung, indem es als Informations- und Steuerungsinstrument die Grundlage für strategische und operative Entscheidungen schafft. Zu den klassischen Kernaufgaben des Controllings zählen die Planung, Kontrolle, Steuerung und Informationsversorgung. Die Planungsaufgabe umfasst dabei die Entwicklung zukunftsorientierter Ziele und Maßnahmen, während die Kontrollfunktion die Überprüfung der Zielerreichung und die Analyse von Abweichungen sicherstellt. Ergänzt wird dies durch die Steuerungsfunktion, die den Abgleich zwischen geplanten und tatsächlichen Unternehmensentwicklungen ermöglicht, um notwendige Korrekturen einzuleiten. Ein weiterer wichtiger Aspekt ist die Bereitstellung relevanter Informationen, die sowohl quantitativ als auch qualitativ auf die Bedürfnisse der Entscheidungsträger abgestimmt sind. Diese Aufgaben bilden gemeinsam einen Kreislauf, der kontinuierlich die Effektivität und Effizienz der Unternehmensaktivitäten überwacht und optimiert.

Mit der digitalen Transformation erfährt nicht nur das Management, sondern auch das Controlling einen weitreichenden Wandel, welcher sowohl Herausforderungen als auch Chancen mit sich bringt. Technologien wie Big Data und KI erweitern die traditionellen Aufgabenfelder des Controllings vgl. (Weißenberger & Bauch, 2019, p. 708f.). Die Einführung moderner Enterprise-Resource-Planning-Systeme (ERP-Systeme) ermöglicht eine Integration von Daten aus verschiedenen Unternehmensbereichen und erlaubt Echtzeitanalysen, die zuvor manuell und zeitaufwendig durchgeführt werden mussten vgl. (Weißenberger & Marrocco, 2022, p. 42). Weiterhin verändert die Digitalisierung grundlegend die Geschäftsprozesse, indem sie zum Großteil standardisiert und automatisiert werden. Dabei werden erhebliche Mengen an Daten freigesetzt und gespeichert, welche für wertschöpfende Aktivitäten genutzt werden können vgl. (Keimer & Egle, 2020, p. 2). Das Controlling ist durch die Digitalisierung in zweifacher Hinsicht betroffen: Zum einen übernimmt es die Aufgabe, den digitalen Wandel im Unternehmen zu steuern (Controlling der Digitalisierung), zum anderen muss es seine eigenen Prozesse an die Erfordernisse der Digitalisierung anpassen (Digitalisierung im Controlling) vgl. (Keimer & Egle, 2020, p. 2), wobei letzteres von übergeordneter Relevanz für diese Arbeit ist.

Ein spezifisches Beispiel für die Chancen der Digitalisierung im Controlling ist die Einführung agiler Methoden und digitaler Plattformen, wie im Fall von Zalando. Zalando hat durch die digitale Transformation bedeutende Fortschritte im Investment-Controlling erzielt, insbesondere durch die Einführung eines Investment-Boardrooms und einer Investment-App vgl. (Engelbers & Moreira, 2020, p. 287). Diese Initiativen dienen dazu, finanzielle Investitionen effizienter zu steuern und gleichzeitig die Prozessabläufe

zu optimieren. Die digitalen Lösungen schaffen Transparenz, die es ermöglicht, Risiken frühzeitig zu erkennen, auf Marktveränderungen flexibel zu reagieren und die Unternehmensstrategie gezielt anzupassen. Darüber hinaus tragen sie zur Beschleunigung und Verbesserung der Qualität von Investitionsentscheidungen bei, während die Arbeitsprozesse durch die digitale Automatisierung schlanker und ressourcenschonender gestaltet wird. Insgesamt fördern diese Projekte die Effektivität des Managements und steigern den langfristigen Unternehmenserfolg vgl. (Engelbers & Moreira, 2020, p. 300).

Nach (Schäffer & Weber, 2016a) wird das Controlling durch die digitale Transformation vor fundamentale Herausforderungen gestellt, die eine tiefgreifende Anpassung von Paradigmen, Aufgaben und Werkzeugen erfordern und im Folgenden erläutert werden. Zunächst führen die zunehmende Automatisierung und Digitalisierung nicht nur zu einer Standardisierung von Prozessen, sondern verlangen auch ein neues Verständnis von Datenmanagement. Damit stehen Controller vor der Aufgabe, nicht nur finanzielle, sondern auch nicht-finanzielle Datenquellen effektiv zu verwalten, um eine fehlerfreie und konsistente Datenbasis sicherzustellen. Diese bildet die Grundlage für bspw. moderne Analyseverfahren wie *predictive analytics*, die datengetriebene Prognosen ermöglichen und neue Einblicke für das Management liefern können vgl. (Schäffer & Weber, 2016a, p. 9).

Gleichzeitig verändert sich die Rolle des Controllers als Gatekeeper von Informationen, da Manager zunehmend in der Lage sind, durch den Einsatz von Self-Controlling-Tools eigenständig auf Steuerungsinformationen zuzugreifen, was den Controller dazu zwingt, sich verstärkt auf die Gestaltung und Optimierung dieser Prozesse zu konzentrieren vgl. (Schäffer & Weber, 2016a, p. 10). Eine weitere Herausforderung besteht darin, dass Controller durch die Digitalisierung zunehmend mit strategischer und operativer Unsicherheit umgehen müssen vgl. (Schäffer & Weber, 2016a, p. 12), weshalb die Unternehmenssteuerung agiler und damit sowohl flexibler, als auch effizienter gestaltet werden muss. Insbesondere digitale Wertschöpfungsketten, wie sie bei Plattformunternehmen wie Zalando oder Amazon zu finden sind, verlangen nach Steuerungsmechanismen, die sich durch eine höhere Anpassungsfähigkeit, optimierte Prozesse und eine stärkere Verzahnung von Daten und Abläufen auszeichnen. Diese Anforderungen können nach den Autoren durch den Einsatz von Automatisierungstechnologien und intelligenten Datenanalysen erfüllt werden, die eine präzisere und schnellere Entscheidungsfindung ermöglichen.

Der intensive Gebrauch von Big Data erfordert die Entwicklung entsprechender analytischer Fähigkeiten im Controlling. Controller müssen sich von klassischen Small-Data-Ansätzen lösen und Kompetenzen ihres Unternehmens in Statistik sowie Informationstechnologie (IT) ausbauen, um datenbasierte Prozesse kritisch zu begleiten und strategische Entscheidungen zu unterstützen. Gleichzeitig stellt sich die Frage, wie Data-Science organisatorisch eingebunden wird und welche Rolle das Controlling dabei übernimmt vgl. (Schäffer & Weber, 2016a, p. 14f).. Infolgedessen gewinnen neben den klassischen Fähigkeiten in Zahlen- und Methodenkompetenz Kenntnisse in Statistik, IT und Datenanalyse sowie auch soziale und kommunikative Fähigkeiten an Bedeutung. Controller müssen sich einem Paradigmenwechsel zu Denkweisen unterziehen, die eine höhere Flexibilität und Risikobereitschaft erfordern vgl. (Schäffer & Weber, 2016a, p. 16). Dazu gehört der Übergang von traditionellen Jahresplanungen zu dynamischen Projektansätzen sowie die Etablierung einer Kultur, die von Experimentierfreude und kontinuierlichem Lernen geprägt ist. Diese Veränderungen stellen nicht nur die Methoden, sondern auch die grundlegende Steuerungslogik des Controllings infrage vgl. (Schäffer & Weber, 2016a, p. 15f)..

Schließlich wird die Effizienzsteigerung im Controlling zu einem entscheidenden Ziel. Automatisierte Prozesse reduzieren die traditionellen Tätigkeiten besonders im operativen Controlling erheblich. Dadurch entsteht Raum für eine strategischere Neuausrichtung, insbesondere in der Arbeit als Business-Partner, was laut (Schäffer & Weber, 2016a, p. 13) für Controller die einzige Möglichkeit ist, ihre Relevanz und Wirksamkeit in einem zunehmend digitalen und dynamischen Umfeld zu sichern. Konkret heißt das, dass der Controller nun verstärkt als kritischer Berater der Unternehmensleitung agiert, dessen Hauptaufgabe darin besteht, datenbasierte Einblicke zu liefern und strategische Entscheidungen aktiv zu unterstützen vgl. (Keimer & Egle, 2020, p. 202). Die Kompetenzen des Business-Partners ermöglichen es, nicht nur die Qualität der Daten und Analysen zu bewerten, sondern auch komplexe Probleme zu durchdringen und strategisch sinnvolle Handlungsempfehlungen abzuleiten vgl. (Keimer und Egle, 2020, p. V; Weißenberger und Bauch, 2019, p. 708f.).

Darüber hinaus wird die Wichtigkeit der Empathie, des ethischen Urteilsvermögens und der interdisziplinären Kommunikation betont, da ein Controller heute in der Lage sein muss, die Perspektiven unterschiedlicher Abteilungen zu verstehen und in die strategische Planung zu integrieren vgl. (Keimer und Egle, 2020, p. VI; Möller et al., 2017, p. 67). Dies ist entscheidend, um Vertrauen und Akzeptanz innerhalb der Organisation aufzubauen und die Umsetzung von Maßnahmen zu erleichtern. Besonders in der Rolle des Business-Partners kommt es darauf an, eine Brücke zwischen den technischen Analysen und der praktischen Umsetzung in der Unternehmensstrategie zu schlagen vgl. (Schäffer & Weber, 2016a, p. 8f.). (Keimer & Egle, 2020, p. V–VI) betonen, dass Controller in ihrer neuen aktiven Rolle als Business Partner sicherstellen müssen, digitale Prozesse verantwortungsvoll und zukunftsorientiert zu begleiten. Dies umfasst sowohl die strategische Ausrichtung auf langfristige Unternehmensziele als auch die Integration sozialer, ökologischer und ethischer Aspekte in die Unternehmenssteuerung.

Zusammengefasst lässt sich das digitale Controlling in fünf zentrale Domänen unterteilen, die den aktuellen Wandel durch die Digitalisierung prägen, welche im folgenden Kapitel näher erläutert werden: Daten, Technologien, Prozesse, Methoden und Kompetenzen vgl. (Keimer & Egle, 2020,

p. 7–14). Diese fünf Domänen zeigen, dass die Digitalisierung das Controlling in diesen Bereichen tiefgreifend verändert und neue Anforderungen entstehen lässt, bei deren Bewältigung CDR eine zentrale Rolle spielt. In den folgenden Kapiteln wird daher untersucht, wie die Digitalisierung die einzelnen Domänen des Controllings beeinflusst und welche spezifischen Anforderungen sich daraus im Kontext der CDR ergeben. Nur wenn klar verstanden wird, welche Bereiche des Controllings durch die Digitalisierung betroffen sind und in welchem Umfang CDR dabei eine Rolle spielt, können fundierte Ansätze und Strategien für die erfolgreiche Implementierung entwickelt werden. Dieses Verständnis bildet die Grundlage, um CDR gezielt in die Prozesse des Controllings zu integrieren und sowohl strategisch als auch operativ umzusetzen.

3 Einfluss von CDR auf die Domänen des digitalen Controllings nach (Keimer & Egle, 2020)

3.1 Daten: Data Governance, Datenschutz und Datenethik

Da Daten eine fundamentale Grundlage für das digitale Controlling bilden, ist es essenziell, im Kontext der CDR einen verantwortungsvollen Umgang mit diesen sicherzustellen. Dies erfordert insbesondere ein effektives Datenmanagement und eine klar definierte Data-Governance, um die Verlässlichkeit, Aktualität und Konsistenz der Daten zu gewährleisten vgl. (Keimer & Egle, 2020, p. 7f.). Dabei entsteht ein Zielkonflikt zwischen der Frage, wie viele und welche Daten für das Controlling ausreichend und relevant sind, und der Notwendigkeit, diese Daten im Hinblick auf Datenschutz und ethische Standards zu schützen. Je umfangreicher die verfügbaren Datenmengen werden, desto schwieriger gestaltet sich die Gewährleistung eines effektiven Datenschutzes. Im Einklang mit dem Grundsatz der Datensparsamkeit vgl. (Thorun, 2018, p. 182) sollten ferner lediglich für das Controlling relevante Daten gesammelt und gespeichert werden. Um das Controlling als *single point of truth* (SPOT) zu erhalten, ist aber eine zentrale Integration aller relevanten Daten unabdingbar vgl. (Schäffer & Weber, 2016a, p. 9f.). Dies verhindert eine fragmentierte Datenlandschaft, wodurch konsistente Entscheidungsgrundlagen möglich sind vgl. (Keimer & Egle, 2020, p. 6). Gleichzeitig ermöglicht der technologische Fortschritt eine Dezentralisierung von Datenzugriffen und Analyseaufgaben, wodurch Self-Service-Ansätze gefördert werden (Möller et al., 2017, p. 65). Der Schlüssel liegt in der Balance zwischen einer zentralen Datenbasis, die für Datenqualität und Konsistenz sorgt, und dezentralen Nutzungsmodellen, die Agilität und Effizienz in der Datenanalyse gewährleisten.

CDR beeinflusst die Speicherung und Nutzung von Daten in zweifacher Hinsicht: Zum einen wirken sich datenschutzrechtliche Normen wie das Bundesdatenschutzgesetz (BDSG) oder die Datenschutzgrundverordnung (DSGVO) direkt auf Controllingsysteme aus, indem sie klare rechtlich bindende Vorgaben für die Datenspeicherung und -verwendung schaffen, die den Schutz personenbezogener Informationen sicherstellen sollen vgl. (Ruff, 2024, p. 193; Rödder, 2018, p. 15). Die DSGVO im Speziellen wirkt sich auf den ganzen EU-Bereich aus und konkretisiert Grundsätze für die Verarbeitung personenbezogener Daten[2], die Rechte von betroffenen Personen und nicht zuletzt Pflichten von Verantwortlichen, die für die Datenverarbeitung zuständig sind vgl. (Ruff, 2024, p. 193). Obwohl die DSGVO in erster Linie den Umgang mit personenbezogenen Daten regelt und daher auf den ersten Blick weniger relevant für das Controlling scheint, da dieses primär mit Finanz- und Unternehmensdaten arbeitet, gibt es zahlreiche Berührungspunkte zu personenbezogenen Daten. So können bspw. Mitarbeiterdaten im Rahmen der Kostenrechnung, Gehaltsanalysen oder für Personalkennzahlen verarbeitet werden[3]. Auch Kundendaten können für Vertriebsanalysen oder Umsatzbewertungen herangezogen werden, ebenso wie personenbezogene Informationen von Geschäftspartnern wie z.B. Lieferanten.

Insbesondere die Grundsätze der Datenminimierung (Art. 5 Abs. 1 Buchst. c) DSGVO) und Zweckbindung (Art. 5 Abs. 1 Buchst. b) DSGVO) erfordern, dass nur die Daten verarbeitet werden, die für den jeweiligen Zweck tatsächlich notwendig sind. Das Controlling muss also sicherstellen, dass personenbezogene Daten nur in dem Umfang verwendet werden, der den rechtlichen Vorgaben entspricht und unnötige Datenverarbeitungen vermeiden. Zudem fordert die DSGVO Transparenz (Art. 5 Abs. 1 Buchst. a) DSGVO) und Rechenschaftspflicht (Art. 5 Abs. 2 DSGVO), was bedeutet, dass die Verarbeitung personenbezogener Daten in Berichten und Analysen dokumentiert und nachvollziehbar sein muss. Dazu gehört, dass offengelegt wird, welche Daten genutzt werden und zu welchem Zweck. Darüber hinaus sind Vorgaben zur Speicherbegrenzung (Art. 5 Abs. 1 Buchst. e) DSGVO) zu beachten, sodass personenbezogene Daten nur so lange gespeichert werden dürfen, wie sie für den jeweiligen Zweck erforderlich sind.

Zum anderen greift CDR auch ethische Grundprinzipien auf, die über die gesetzlichen Anforderungen hinausgehen und darauf abzielen, den verantwortungsvollen und fairen Umgang mit Daten zu fördern. Die CDR knüpft also auch da an, wo die DSGVO und andere Gesetze an ihre Grenzen kommen, wie z.B. der Bezug zu Menschenrechten vgl. (Dörr, 2020, p. 78). Ein transnationaler Bezugspunkt für diese ethischen Grundsätze sind die UN-Leitprinzipien für Wirtschaft und Menschenrechte, welche die Verantwortung von Unternehmen zur Wahrung der Menschenrechte in ihren Geschäftsprozessen betonen (Noti et al., 2020, p. 17). Dies schließt die Pflicht ein, negative Menschenrechtsauswirkungen zu vermeiden und Abhilfe zu schaffen, wenn diese auftreten vgl. (OHCHR, 2011, p. 14). Für das Controlling be-

[2] GemäßArt. 4 Nr. 1 DSGVO handelt es sich bei personenbezogenen Daten um „*alle Informationen, die sich auf eine identifizierbare natürliche Person beziehen*“.

[3] Vgl. https://www.hrworks.de/lexikon/hr-kennzahlen/, letzter Zugriff am 25.01.2025.

deutet dies, dass Prozesse und Analysen so gestaltet werden müssen, dass sie keine unethischen oder menschenrechtswidrigen Praktiken fördern. Auch der CDR-Kodex, welcher sich an ethischen Wertevorstellungen im digitalen Raum orientiert, definiert neun Leitprinzipien, bestehend aus der Befolgung gesellschaftlicher Grundwerte (bspw. Demokratie, Gleichbehandlungsgrundsatz), Menschenzentrierung, Schaffung eines Mehrwerts, Schadensvermeidung, Autonomie (der Verbraucher), Fairness, Transparenz, Verantwortlichkeit und Nachhaltigkeit vgl. (Corporate Digital Responsibility-Initiative, 2021, p. 3; Ruff, 2024, p. 178). Digitale Ethik und CDR bezeichnen somit nicht nur abstrakte Ideen, sondern liefern konkrete Wertansichten, die ein Unternehmen nutzen kann, um seine digitale Verantwortung sowohl theoretisch zu rechtfertigen als auch praktisch umzusetzen vgl. (Ruff, 2024, p. 179). Die Prinzipien schaffen eine Orientierung, die das Unternehmen glaubwürdig macht und gleichzeitig klare Handlungsanweisungen bietet.

3.2 Technologien und Prozesse: Verantwortungsvoller Einsatz von Künstlicher Intelligenz und Big Data

Technologien bilden die Grundlage für die Erhebung, Verarbeitung und Analyse betriebswirtschaftlicher Daten. (Keimer & Egle, 2020, p. 9) betonen, dass das Controlling mithilfe moderner ERP-Systeme, welche mit Business-Intelligence- (BI) und Business-Analytics- (BA) Lösungen ergänzt werden, in der Lage sein muss, Daten zu sammeln und zugänglich zu machen, sodass sie für strategische Entscheidungen nutzbar sind. Es handelt sich bei BI- bzw. BA-Lösungen um Anwendungen zur Sammlung, Auswertung und Analyse von Echtzeitdaten aus allen Bereichen des Unternehmens mit dem Ziel, Datenqualität und den SPOT im Unternehmen zu gewährleisten vgl. (Bharadiya, 2023, p. 16; Hecht und Scherrer, 2020, p. 86). Durch eine gute Datenbasis, welche mithilfe innovativer Technologien verarbeitet wird, kann in der Domäne Prozesse ein hoher Automatisierungsgrad sowie eine hohe Effizienz ermöglicht werden vgl. (Keimer & Egle, 2020, p. 10f)..

KI stellt in diesem Kontext eine transformative Ergänzung der BI-Technologien dar. Sie erweitert klassische BI-Tools durch die Fähigkeit, große und unstrukturierte Datenmengen (Big Data) zu analysieren und daraus Muster sowie Vorhersagen abzuleiten vgl. (Dörr, 2020, p. 17). Die Integration von KI in das Controlling eröffnet zahlreiche Chancen, da KI-gestützte Algorithmen Routineaufgaben wie die Datenbereinigung und -analyse automatisieren können, wodurch Controller mehr Zeit für strategische Aufgaben gewinnen (Hecht & Scherrer, 2020, p. 98).

Gleichzeitig bringt der Einsatz von KI auch erhebliche Herausforderungen mit sich. Eine der zentralen Schwierigkeiten besteht in der mangelnden Transparenz vieler KI-Systeme, da diese oft als sogenannte Black-Boxes agieren vgl. (Floridi et al., 2018, p. 692). Die Entscheidungsprozesse solcher Modelle sind für Anwender schwer nachvollziehbar, was zu einem Konflikt mit ethischen Anforderungen wie Fairness und Verantwortlichkeit führen kann.[4] Hinzu kommt das Risiko von Verzerrungen (*biases*), die durch unausgewogene Trainingsdaten, die Wahl statistischer Methoden, sowie den Einfluss latenter Variablen oder zufällig zusammenhängende Korrelationen entstehen können und diskriminierende Entscheidungen begünstigen vgl. (Kunz und Wirtz, 2024, p. 36; Weißenberger und Bauch, 2019, p. 711f). Der Einsatz von KI bringt schließlich auch Bedenken hinsichtlich der Verantwortung gegenüber Mitarbeitern, einschließlich der Sicherung von Arbeitsplätzen, der Förderung von Loyalität und KI-Kompetenzen, der Wahrung der Privatsphäre und des Schutzes der persönlichen Daten mit sich vgl. (Dörr, 2020, p. 26).

All diese Bedenken führen dazu, dass der Einsatz von KI-Systemen für Prognosen und Handlungsempfehlungen vor allem nur dann sinnvoll sein kann, wenn die zugrunde liegenden Modelle auf fundierten und gut nachvollziehbaren theoretischen Grundlagen basieren, die über den betrachteten Zeitraum hinweg unverändert bleiben. In solchen Fällen überwiegen die Vorteile, die durch das Lernen und die Weiterentwicklung der Algorithmen entstehen, die potenziellen Nachteile, die durch fehlerhafte Entscheidungen verursacht werden könnten vgl. (Weißenberger & Bauch, 2019, p. 713). Um einen verantwortungsvollen Umgang mit KI in Zukunft etablieren zu können und die Akzeptanz innerhalb der Belegschaft zu festigen, werden CDR-bezogene Fragestellungen relevant.

CDR verpflichtet Unternehmen, ethische Prinzipien wie Fairness, Transparenz und Rechenschaftspflicht in den gesamten Lebenszyklus von KI-Anwendungen zu integrieren. Dies bedeutet konkret, dass KI stets darauf ausgerichtet sein sollte, einerseits nur Gutes zu tun (*beneficence*) und andererseits keinen Schaden zu verursachen (*non-maleficence*) vgl. (Floridi et al., 2018, p. 696). *Beneficence* umfasst die Förderung des menschlichen Wohlergehens, die Achtung der Würde aller Menschen sowie die Wahrung der planetaren Nachhaltigkeit. Soziale und ökologische Variablen sollten somit neben den ökonomischen Variablen in KI-gestützte unternehmerische Entscheidungen miteinbezogen werden. *Non-maleficence* hingegen legt den Schwerpunkt auf die Vermeidung von Schäden, indem potenzielle Risiken wie Datenschutzverletzungen, Missbrauch oder unvorhergesehene schädliche Auswirkungen von KI-Systemen minimiert werden (Floridi et al., 2018, p. 697).

Die EU reguliert den Einsatz von KI durch die am 1. August 2024 in Kraft getretene Verordnung 2024/1689, welche zukunftsweisende Anforderungen an Transparenz und Sicherheit von KI-Systemen stellt.[5] Insbesondere KI-Systeme mit hohem Risiko, wie sie im Controlling eingesetzt werden könnten, unterliegen demnach einer Reihe von Verpflichtungen, bevor sie frei in Umlauf kommen, bspw. einer umfassenden Risikoanalyse, einer lückenlo-

[4] Vgl. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, letzter Zugriff am 20.01.2025.

[5] Vgl. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, letzter Zugriff am 20.01.2025.

sen Dokumentation der Aktivitäten und Konformität vgl. (Datenschutz-Grundverordnung (DSGVO), 2016, p. 3).

Zukünftig könnte *explainable AI* (XAI) ein Ansatz darstellen, um die immer komplexer werdenden KI-gestützten Analysen und Vorhersagen nachvollziehbar und transparent zu gestalten. Es handelt sich bei XAI um Prozesse und Methoden, die es ermöglichen, die Entscheidungslogik von KI-Modellen offenzulegen, was besonders für Controller wichtig ist, da sie diese Ergebnisse für strategische Entscheidungen nutzen vgl. (Rai, 2020, p. 137f). Durch die Nachvollziehbarkeit der zugrunde liegenden Prozesse wird die Black-Box-Problematik reduziert, wodurch Datenanalysen effektiver geprüft und besser in die Unternehmensstrategie integriert werden können, während ein hohe Genauigkeit erhalten bleibt (Ali et al., 2023, p. 2f.). Diese Maßnahmen tragen dazu bei, Vertrauen in KI-Systeme zu schaffen, deren Akzeptanz zu fördern und sicherzustellen, dass sie verantwortungsvoll eingesetzt werden.

Der Einbezug von Big Data in KI-gestützte Entscheidungen und damit der Einbezug großer Datenmengen ermöglicht eine revolutionäre Optimierung von Prozessen, präzisere Analysen und die Identifikation von Trends und Risiken in Echtzeit, wodurch Manager fundiertere Entscheidungen treffen können, die weniger auf Intuition basieren vgl. (Dörr, 2020, p. 17; McAfee und Brynjolfsson, 2012, p. 5). Dabei haben Investitionen in Big Data einen positiven messbaren Einfluss auf den Unternehmenserfolg, wenn vor allem kulturelle und organisatorische Barrieren überwunden werden können vgl. (Davenport & Bean, 2017, p. 6). Gleichzeitig stellt Big Data Unternehmen vor ethische und ökologische Herausforderungen. Zum einen müssen Datenschutzrichtlinien wie die DSGVO beachtet werden, um die Rechte der Betroffenen zu schützen vgl. (Ruff, 2024, p. 193). Zum anderen ist die lokale Speicherung und Verwaltung erheblicher Datenmengen sowie eine KI-gestützte Auswertung ein energieintensives Unterfangen, da eine ausreichend leistungsfähige Hardware benötigt wird vgl. (Gotsch et al., 2022, p. 260; Manhart, 2024).

CDR fordert in diesem Zusammenhang, dass Unternehmen klare Richtlinien für den Umgang mit Big Data und KI entwickeln, welche die Prinzipien des Datenschutzes, der Transparenz und der ökologischen Nachhaltigkeit adressieren, um das Vertrauen der Stakeholder zu stärken.

3.3 Methoden und Kompetenzen: Analytics und Wandel der Controller-Rolle

Moderne Analysemethoden wie *diagnostic*, *predictive* und *prescriptive analytics* ermöglichen es, unstrukturierte Big Data über den bisherigen Informationsgehalt hinaus zu interpretieren, darzustellen und für strategische Entscheidungen nutzbar zu machen vgl. (Weißenberger & Bauch, 2019, p. 710). Dies erfordert nicht nur technisches und methodisches Know-How, sondern auch ein Umdenken in den Rollen und Kompetenzen von Controllern. Es stellt sich die Frage, wie Controller unter Beachtung von CDR durch den Einsatz dieser Analysemethoden eine aktivere Rolle als Business Partner einnehmen können und den Wandel ins Digital Controlling meistern. Ziel ist es, die Entwicklung des Controllers als strategischen Gestalter und Treiber digitaler Transformation im Unternehmen darzustellen.

Die Beherrschung von Analysemethoden wie *descriptive analytics* (Beschreibung von Ist-Zuständen), *diagnostic analytics* (Analyse von Ursachen und Zusammenhängen), *predictive analytics* (Prognosen über zukünftige Entwicklungen und Trends) und *prescriptive analytics* (Empfehlungen für Handlungsoptionen) ermöglichen es Controllern, nachvollziehbare datenbasierte Steuerungsentscheidungen zu treffen vgl. (Keimer & Egle, 2020, p. 11). Diese Methoden erfordern den Einsatz spezialisierter Software, leistungsstarker IT-Infrastrukturen und fortgeschrittenes methodisches Wissen. Insbesondere p*redictive* und *prescriptive analytics*, welche im Controlling zukünftig von hoher Relevanz sind, helfen dabei, Stärken und Schwächen eines Geschäftsbereiches rechtzeitig aufzuzeigen und Handlungsempfehlungen vorzubereiten vgl. (Gerig, 2020, p. 235). Zusätzlich kann die Ableitung von Unternehmensentscheidungen unter Verwendung von BA-Systemen teilweise menschliche Schwächen im Entscheidungsverhalten, wie z.B. *biases* reduzieren vgl. (Weißenberger & Bauch, 2019, p. 714). Nach (Caviezel, 2020, p. 119) kann sich das Controlling durch die Nutzung von BA-Systemen vermehrt mit *predictive analytics* auseinandersetzen, statt mit vergangenheitsbezogenen Soll-Ist-Vergleichen und als Business-Partner des Managements bei der Entscheidungsfindung mitwirken. CDR beeinflusst diese Domäne insofern, dass die Nutzung der Analysemethoden im Einklang mit gesellschaftlichen Erwartungen stehen, indem z.B. sichergestellt wird, dass diese Methoden nicht unter Einfluss diskriminierender Algorithmen Handlungsempfehlungen ableiten.

Wie bereits in Kapitel 2.2 dargestellt, erfordert der digitale Wandel im Unternehmen ein Umdenken der Controllerrolle und damit verbunden eine Verschiebung der Kompetenzanforderungen für Controller. In der letzten Domäne „Kompetenzen“ geht es also darum, Anforderungen an den Controller zu beschreiben, der für die Digitalisierung des Controllings und für das Controlling der Digitalisierung verantwortlich ist vgl. (Keimer & Egle, 2020, p. 12). Denn nur durch die Kompetenzen des digitalen Controllers gelingt es dem Unternehmen, die Digitalisierung in allen zuvor analysierten Domänen voranzutreiben und dabei CDR als Rahmenbedingung zu etablieren. (Keimer & Egle, 2020, p. 13f.) priorisieren hierfür die folgenden Kompetenzen „Fachwissen“, welches die Grundlage für die Aufgabenbewältigung darstellt, „Business“, welches u.a. Markt- und Branchenverständnis sowie Verständnis über Recht, Ethik und Compliance mit sich bringt, „Data-Science“, welches wichtig für die Verwendung von BI-/BA-Systemen ist, „IT-Management“, wozu IT-Governance, -Security und -Architekturen, Datenmanagement, Technologien gehören und „Leistungskultur“, die von bspw. Eigeninitiative, Kommunikationsfähigkeit, Genauigkeit, einer analytischen Denkweise und Belastbarkeit geprägt ist. Ferner besteht Einigkeit darüber, dass der Controller seine neue Rolle als effektive und unverzichtbare Schnittstelle zwischen den Unternehmensbereichen, die Daten beschaffen, auswerten

und für Entscheidungen benutzen, wahrnehmen muss, um den SPOT im Controlling zu erhalten vgl. (Caviezel, 2020, p. 119; Hecht und Scherrer, 2020, p. 98). Voraussetzung für diese führende Rolle ist ein hohes Maß an Know-How, Kommunikationsfähigkeit und Durchsetzungsvermögen vgl. (Schäffer & Weber, 2016b, p. 497f.).

Die von (Keimer & Egle, 2020, p. 13f.) definierten Kompetenzdimensionen – Fachwissen, Business, Data-Science, IT-Management und Leistungskultur – bieten eine Grundlage, um die Herausforderungen der Digitalisierung zu meistern. Im Kontext von CDR gewinnen diese Dimensionen jedoch zusätzliche Bedeutung, in dem die Anforderungen erweitert werden (Abb. 2). Im Bereich Fachwissen werden nicht nur Kenntnisse über digitale Nachhaltigkeitskennzahlen und ein tiefgreifendes Verständnis von Controllingprozessen relevant, sondern auch in Datenschutz und ethischen Standards, um Risiken wie algorithmische Diskriminierung zu vermeiden. Die Dimension Business verlangt ein erweitertes Verständnis von Compliance, Nachhaltigkeitsstrategien und digitalen Geschäftsmodellen, um datengetriebene Entscheidungen verantwortungsvoll und branchenspezifisch zu gestalten. Die Kompetenz in Data Science muss über die bloße Nutzung von Analysetools hinausgehen und die Fähigkeit umfassen, Transparenz und Nachvollziehbarkeit in Algorithmen sicherzustellen und frühzeitig *biases* zu erkennen und zu korrigieren. Ähnlich fordert IT-Management den Einsatz nachhaltiger Technologien und sicherer Dateninfrastrukturen, die den Prinzipien der CDR entsprechen und regelmäßig danach geprüft werden sollten. Schließlich verlangt die Dimension Leistungskultur nicht nur Eigeninitiative und Teamfähigkeit, sondern auch die Fähigkeit, ethische Herausforderungen eigenverantwortend und interdisziplinär anzusprechen, sowie CDR als integralen Bestandteil der Unternehmenskultur zu verankern. Demzufolge sollte CDR nicht als eigenständige Dimension betrachtet werden, sondern als übergreifende Grundlage, die alle Kompetenzbereiche des digitalen Controllers durchzieht. Um dies zu verdeutlichen, werden die in Abb. 2 genannten Dimensionen um einen äußeren Ring ergänzt, der die CDR-Kompetenz als Rahmenbedingung darstellt. Dieser Ansatz unterstreicht, dass CDR nicht isoliert, sondern integriert betrachtet werden muss, da es die Anforderungen an Fachwissen, Business, Data-Science, IT-Management und Leistungskultur gleichermaßen beeinflusst.

Aus den durch die Digitalisierung gestiegenen Kompetenzanforderungen an das Controlling wird ersichtlich, dass ein einzelner Controller nicht alle diese Dimensionen und Facetten hinreichend für das Controlling der Digitalisierung im Unternehmen mitbringen kann. In großen Unternehmen ist es daher essenziell, dass die unterschiedlichen Kompetenzdimensionen im Controlling-Team, idealerweise bestehend aus digitalen Controllern, technischen Experten, Reportern und Navigatoren vollständig abgedeckt werden vgl. (Trachsel & Bitterli, 2020, p. 208). Für kleine und mittlere Unternehmen (KMU) stellt dies jedoch eine besondere Herausforderung dar, da sich KMU in der Regel keine großen und spezialisierten Controlling-Teams leisten können – sie sollten diese

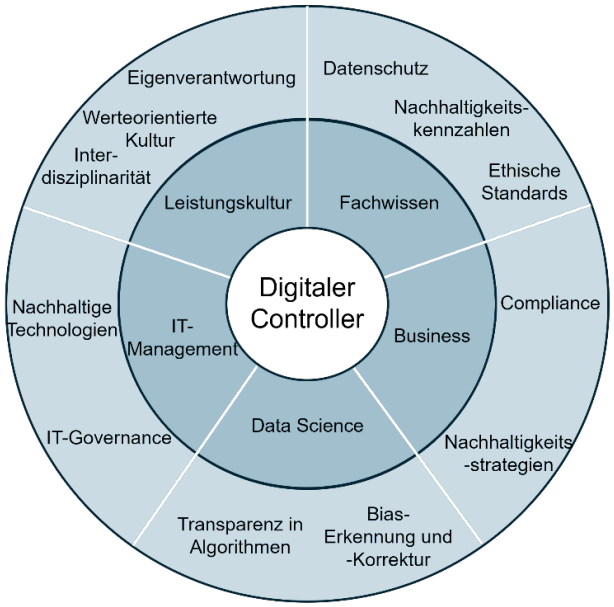

Abbildung 2: CDR-Erweiterung der Kompetenzen des digitalen Controllers (in Anlehnung an Keimer/Egle 2018, S. 51)

Dimensionen priorisieren und bei Neueinstellungen gezielt auf CDR-relevante Kompetenzen achten. Dies erfordert eine sorgfältige Abwägung zwischen wirtschaftlichen Zwängen und der Notwendigkeit, die ethischen und gesellschaftlichen Anforderungen der digitalen Transformation zu erfüllen. Alternativ schlagen (Trachsel & Bitterli, 2020, p. 208) den Einsatz standardisierter Software oder Outsourcing der Controllingfunktionen für KMU vor.

Zusammenfassend beeinflusst CDR als übergreifendes Rahmenkonzept alle Domänen des digitalen Controllings. Insbesondere die Domäne der Kompetenzen wird durch CDR nachhaltig geprägt, da die ethischen, nachhaltigen und rechtlichen Anforderungen die Kompetenzanforderungen an digitale Controller maßgeblich erweitern. CDR wirkt also nicht als isolierte Dimension, sondern als Grundprinzip, das alle Kompetenzbereiche durchdringt. Diese Integration von CDR als verbindendes Element hebt die strategische Bedeutung von CDR im digitalen Controlling hervor und stärkt die Rolle des Controllers als ethischen und nachhaltigen Gestalter der digitalen Transformation. Um jedoch sicherzustellen, dass CDR nicht nur theoretisch berücksichtigt, sondern auch praktisch umgesetzt wird, bedarf es konkreter Strategien und Ansätze. Das folgende Kapitel widmet sich somit der Frage, welche Herausforderungen bei der Umsetzung entstehen und wie CDR effektiv im digitalen Controlling gestaltet und implementiert werden kann. Dabei werden praxisorientierte Maßnahmen und Instrumente vorgestellt, die Unternehmen dabei unterstützen, CDR systematisch in ihre Controllingprozesse zu integrieren.

4 Gestaltung und Umsetzung von CDR im Digitalen Controlling

4.1 Implementierungsstrategien von CDR im Controlling

Die Umsetzung von CDR stellt Unternehmen vor erhebliche Herausforderungen. Laut einer Deloitte-Studie geben 43 % der befragten Unternehmen an, dass ihnen die nötigen Kompetenzen für die Implementierung von CDR fehlen und mehr als 30 %, dass es an konkreten Konzepten mangelt, wie CDR zielführend und nachhaltig umgesetzt werden kann vgl. (Andersen, 2022, p. 13). Wie bereits im vorherigen Kapitel dargestellt, betrifft CDR zahlreiche Domänen im digitalen Controlling und erfordert daher eine strukturierte und koordinierte Herangehensweise. Ohne ein durchdachtes Konzept besteht die Gefahr, dass die vielfältigen Anforderungen nicht angemessen adressiert werden, was die Entstehung von Insellösungen begünstigt vgl. (Keimer & Egle, 2020, p. 6).

Es bedarf daher praxisorientierter und strukturierter Ansätze, die sowohl die strategische als auch die operative Integration von CDR ermöglichen. Der CDR-Kompass von (Weißenberger & Marrocco, 2022, p. 45) und das iterative 6-Schritte-Modell von (Dörr, 2020, p. 53) stellen in der wissenschaftlichen Literatur zwei Ansätze zur Implementierung von CDR dar. Im Folgenden werden beide Ansätze vorgestellt und ihre Gemeinsamkeiten sowie Unterschiede analysiert und in den Kontext des Controllings gesetzt.

Die folgenden Ausführungen zum CDR-Kompass beziehen sich auf die Handlungsfelder der digitalen Verantwortung von (Weißenberger & Marrocco, 2022, p. 46–48). Der CDR-Kompass (Abb. 3) bietet ein Rahmenkonzept, welches die Integration von CDR auf strategischer und operativer Ebene gleichermaßen ermöglicht. Sein Ansatz identifiziert vier zentrale Handlungsfelder (finanziell, ökologisch, sozial und Governance-bezogen), die als Orientierungspunkte für eine verantwortungsvolle Umsetzung von CDR in allen Handlungsfeldern dienen. Diese Handlungsfelder bestehen zunächst aus dem primären Verständnis von CDR, welches strukturell im Compliance-Management verankert werden muss. Hierbei geht es um die Befolgung rechtlicher Pflichten als Mindestanforderung von CDR, was analog zur rechtlichen Verantwortung der CSR-Pyramide nach (Carroll, 1991) betrachtet werden kann.

Ferner müssen die vier Handlungsfelder auf der nächsten Handlungsebene als integraler Bestandteil in die operative Unternehmensführung etabliert werden. Auf dieser Handlungsebene muss CDR in einem transaktionalen Verständnis betrachtet werden, das sich auf die tägliche Umsetzung von Maßnahmen innerhalb bestehender Prozesse bezieht. Im digitalen Controlling bedeutet dies, dass die vier Handlungsfelder operativ in Steuerungs- und Kontrollmechanismen integriert werden müssen. Dadurch kann das Controlling bspw. durch den Einsatz nachhaltiger, ethischer, und energieeffizienter Technologien die Umsetzung dieser Handlungsfelder direkt unterstützen und fördern. Durch die Implementierung von CDR in operativen Prozessen des Controllings können Wettbewerbsvorteile erzielt

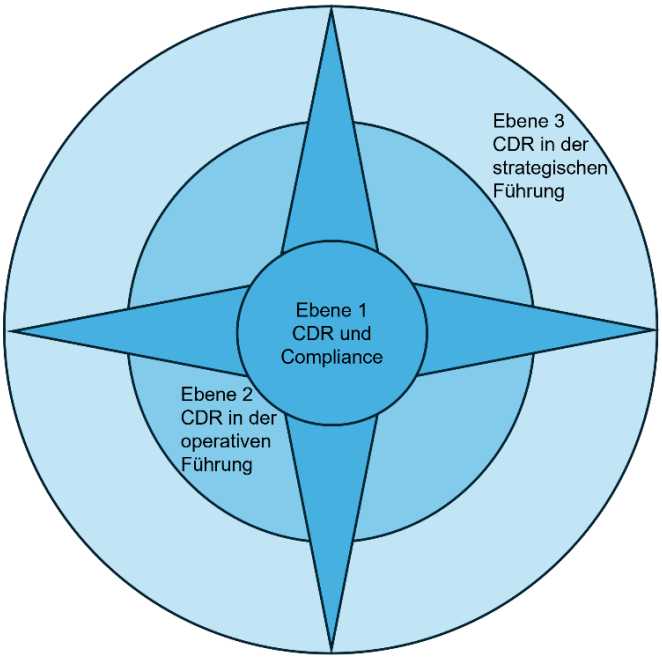

Abbildung 3: CDR-Kompass (Weißenberger & Marrocco, 2022, p. 45)

werden, indem Effizienzpotenziale ausgeschöpft, Risiken minimiert und Innovationsmöglichkeiten besser genutzt werden. Darüber hinaus ermöglicht die Einbindung von CDR auf der nächsten strategischen Handlungsebene der Unternehmensführung eine nachhaltigere Ausrichtung, die langfristige Wettbewerbsvorteile schafft. Insbesondere durch die Schaffung sogenannter Begeisterungseigenschaften, wie etwa Transparenz, ethischer Entscheidungsfindung oder Nachhaltigkeitsberichterstattung, kann das Controlling einen signifikanten Beitrag zur Stakeholder-Bindung leisten. Die Autoren betonen, dass die strategische und operative Integration der vier Handlungsfelder von CDR ein ganzheitliches Verständnis und eine systematische Verankerung in allen Unternehmensbereichen erfordert. Der Kompass dient dabei nicht nur als Orientierungsrahmen, sondern auch als praktisches Werkzeug, um die Verantwortung des Unternehmens entlang klar definierter Handlungsfelder umzusetzen.

Das iterative Modell von (Dörr, 2020, p. 52) bietet einen praxisorientierten Ansatz, um CDR schrittweise im Unternehmen zu etablieren (Abb. 4). Es basiert auf einem Innovationszyklus, der in sechs aufeinander abgestimmten Schritten organisiert ist, welche im Folgenden erläutert werden vgl. (Dörr, 2020, p. 52–54): Der Prozess beginnt mit dem Schritt „Know-How!“, der darauf abzielt, ein grundlegendes Verständnis für die digitale Transformation und deren Auswirkungen auf die Unternehmensverantwortung zu schaffen. Es sollen die grundlegenden technologischen Entwicklungen und deren Implikationen für Wirtschaft und Gesellschaft analysiert werden, um die Notwendigkeit von CDR zu erkennen und ein gemeinsames Verständnis dafür zu entwickeln.

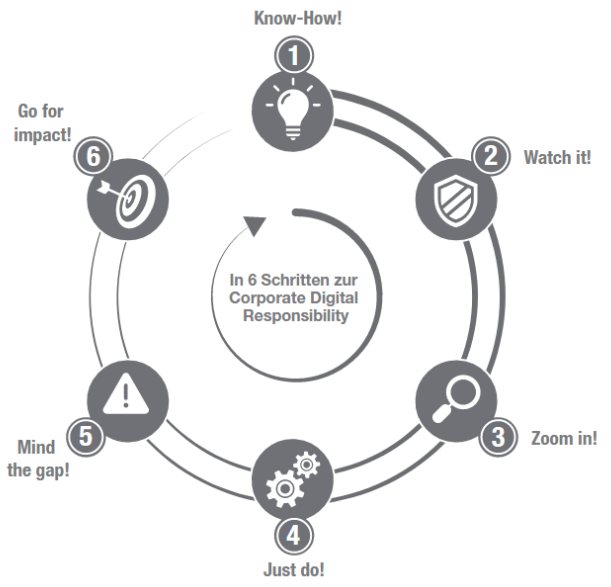

Abbildung 4: Das 6-Schritte-Modell zur Implementierung von CDR (Dörr, 2020, p. 52)

Im zweiten Schritt, „Watch it!“, werden die Chancen und Risiken der Digitalisierung im Hinblick auf Nachhaltigkeit betrachtet. Hierbei wird untersucht, welche positiven Beiträge die Digitalisierung zur gesellschaftlichen und ökologischen Entwicklung leisten kann, während gleichzeitig negative Aspekte der Digitalisierung wie etwa erhöhte Umweltbelastungen oder soziale Ungleichheiten analysiert werden. Dieser Schritt sensibilisiert Unternehmen für die potenziellen Auswirkungen der Digitalisierung und legt den Grundstein für eine nachhaltige Ausrichtung.

Der dritte Schritt, „Zoom in!“ dient dazu, den aktuellen Stand der CDR im Unternehmen zu bestimmen und zu bewerten. Dabei können mithilfe von Modellen wie z.B. *digital responsibility checks* der Reifegrad und die Ganzheitlichkeit bestehender Ansätze überprüft werden. Ziel dieses Schrittes ist es, die spezifischen Chancen und Risiken für das Unternehmen zu identifizieren und eine fundierte Grundlage für die weitere Umsetzung zu schaffen.

Im vierten Schritt, „Just do!“, geht es um die praktische Umsetzung von CDR-Maßnahmen im Unternehmen. Dabei werden strategische und operative Handlungsoptionen entwickelt, die unter anderem bestehende Corporate Responsibility-Instrumente einbeziehen. Dieser Schritt umfasst auch die Anpassung von Stakeholder-Analysen sowie die Entwicklung digitaler Selbstverpflichtungen, um Verantwortung sowohl intern als auch extern sichtbar zu machen. Darüber hinaus wird die Integration von CDR in Innovationsprozesse und die Geschäftsmodellentwicklung gefördert.

Im fünften Schritt, „Mind the gap!“, werden praktische Herausforderungen bei der Umsetzung von CDR thematisiert. Unternehmen sollen potenzielle Hindernisse wie Unsicherheit oder spezifische Barrieren innerhalb der Organisation identifizieren und bewältigen. Hilfreich bei der Orientierung sind dabei Beispiele aus der Unternehmenspraxis, wie CDR bereits vereinzelt erfolgreich implementiert worden ist, was im Controlling zurzeit jedoch eine Rarität darstellt.

Der letzte Schritt, „Go for impact!“, fokussiert sich darauf, die Wirkung von CDR-Maßnahmen zu evaluieren und nachhaltig zu steigern. Dabei wird betont, dass Verantwortung nicht nur in die strategische Unternehmensführung integriert werden muss, sondern auch messbare Ergebnisse erzielen sollte. Unternehmen sollen die Wirkung ihrer digitalen Verantwortung anhand interner und externer Ziele bewerten und dabei das Risiko eines ethischen Theaters vermeiden, durch das CDR-Maßnahmen lediglich symbolisch wirken oder Greenwashing-Vorwürfe hervorrufen. Stattdessen wird eine konsequente Verankerung von CDR im strategischen Management angestrebt, um sowohl den Unternehmenswert als auch den gesellschaftlichen Beitrag zu steigern.

Dörrs Modell hebt hervor, dass die Schritte nicht linear, sondern iterativ zu verstehen sind. Unternehmen können den Prozess flexibel anpassen und bei Bedarf einzelne Schritte erneut durchlaufen, um die CDR-Strategie kontinuierlich zu optimieren vgl. (Dörr, 2020, p. 54).

Beide Ansätze teilen die grundlegende Auffassung, dass CDR eine zentrale Rolle in der Unternehmensführung spielen muss. Sowohl der CDR-Kompass als auch das 6-Schritte-Modell erkennen die strategische Bedeutung von CDR und betonen die Notwendigkeit einer systematischen Integration in operative Prozesse und Strukturen. Beide Ansätze legen zudem Wert auf die Verbindung von Nachhaltigkeit, sozialer Verantwortung und technologischer Innovation.

Die Unterschiede liegen vor allem in der Zielsetzung und Herangehensweise. Der CDR-Kompass bietet ein konzeptionelles Framework, das sich auf die strategische Verankerung von CDR in der Unternehmensführung konzentriert. Es legt weniger Wert auf konkrete operative Maßnahmen, sondern dient als Orientierungshilfe für langfristige Planungen. Im Gegensatz dazu ist Dörrs Ansatz praxisorientierter und liefert konkrete Handlungsschritte und Empfehlungen, die Unternehmen direkt umsetzen können. Der iterative Charakter des Modells ermöglicht eine flexible Anpassung an spezifische Unternehmensbedingungen und macht es besonders geeignet für die operative Umsetzung, welches sich für die Unternehmenspraxis als besonders geeignet erweist. Der CDR-Kompass hingegen ist wertvoll für die strategische Planung und die Entwicklung eines übergeordneten Verständnisses von CDR, aber weniger direkt anwendbar. Eine Kombination beider Ansätze könnte eine ideale Lösung darstellen, bei der der strategische Rahmen des CDR-Kompasses mit den operativen Handlungsschritten des 6-Schritte-Modells ergänzt wird. Der CDR-Kompass lässt sich besonders in zwei zentrale Schritte des iterativen 6-Schritte-Modells von Dörr integrieren, um die strategische und operative Umsetzung von CDR optimal zu unterstützen. Diese Kombination verbindet die klare strategische Ausrichtung des Kompasses mit der flexiblen und praxisnahen Struktur des Dörr-Modells.

Im Schritt „Zoom in!“ des 6-Schritte-Modells, kann der CDR-Kompass als Bewertungsinstrument eingesetzt werden.

Der Kompass bietet eine Möglichkeit, den aktuellen Stand der CDR in Bezug auf die vier Handlungsfelder zu bestimmen, bestehende Strategien und Aktivitäten zu analysieren und Schwachstellen sowie Potenziale zu identifizieren. Die Handlungsfelder ermöglichen eine differenzierte Bewertung der unternehmerischen Chancen und Risiken sowie die Priorisierung von Handlungsfeldern, die für das Unternehmen besonders relevant sind. Somit würde die Einbindung des CDR-Kompasses in diesen Schritt nicht nur die Bewertung detaillierter gestalten, sondern auch sicherstellen, dass sie strategisch auf die langfristigen Unternehmensziele ausgerichtet ist.

Ein weiterer Ansatzpunkt für die Integration des CDR-Kompasses ist der Schritt „Just do!“, der sich auf die strategische und operative Umsetzung von Maßnahmen konzentriert. Hier kann der Kompass als Leitlinie dienen, welche sicherstellt, dass die entwickelten Maßnahmen strategisch fundiert und systematisch auf die vier zentralen Handlungsfelder ausgerichtet sind. Für jedes Handlungsfeld könnten bspw. spezifische Maßnahmen entwickelt werden, wie die Förderung nachhaltiger IT-Infrastrukturen im ökologischen Bereich oder die Sicherstellung diskriminierungsfreier Algorithmen im sozialen Bereich.

Die Kombination beider Ansätze bietet eine starke Synergie, indem sie die strategischen Ziele und die operativen Anforderungen von CDR miteinander verknüpft. Der CDR-Kompass definiert klare Handlungsfelder, die im Rahmen des iterativen Prozesses des Dörr-Modells kontinuierlich überprüft und angepasst werden können. Dies ermöglicht Unternehmen nicht nur eine flexible Anpassung an die dynamischen Bedingungen der Digitalisierung, sondern auch eine langfristige Verankerung von CDR sowohl auf strategischer als auch auf operativer Ebene.

4.2 Herausforderungen und Lösungsansätze bei der Umsetzung

Die Schritte „Just do!“ und „Mind the Gap!“ aus dem 6-Schritte-Modell sind ebenso trivial wie komplex. Um eine erfolgreiche Implementierung von CDR im Controlling durchzusetzen, ist es von besonderer Bedeutung, die organisationsspezifischen Herausforderungen zu kennen. In diesem Kapitel soll es darum gehen, Herausforderungen bei der Umsetzung aufzuzeigen und Lösungsansätze herauszuarbeiten.

In erster Linie sind Manager und Controller damit anvertraut, CDR durchzusetzen, was jedoch zu Problemen führen kann, denn ein theoretisches Konzept allein bedeutet noch nicht, dass es auch praktisch umgesetzt wird, da Hürden wie die Komplexität des jungen Themengebiets oder die mangelnde Messbarkeit hinsichtlich der Erfolgswirksamkeit die Umsetzung von CDR-bezogenen Entscheidungen hindern können vgl. (Dörr, 2020, p. 163; Weißenberger und Marrocco, 2022, p. 48). Die daraus resultierende Steuerungslücke beschreibt die Diskrepanz zwischen strategisch gewünschten Maßnahmen und deren tatsächlicher Implementierung auf operativer Ebene vgl. (Weißenberger & Bauch, 2019, p. 706). Hierbei bietet das Object-of-Control-Framework nach (Merchant & Van der Stede, 2023) eine wertvolle Grundlage, um die Ursachen dieser Lücke zu analysieren, indem es drei zentrale unterscheidbare Barrieren identifiziert: fehlendes Wissen, mangelnde Motivation und persönliche Einschränkungen vgl. (Merchant & Van der Stede, 2023, p. 9). Um diese Hürden zu überwinden, schlägt das Framework einerseits Maßnahmen zur Problemvermeidung wie z.B. Rückdelegation, Automatisierung oder Zentralisierung vor, andererseits sogenannte *management control systems*, welche sich in die Kategorien Ergebnissteuerung, Prozesssteuerung und soziale Steuerung untergliedern lassen vgl. (Merchant & Van der Stede, 2023, p. 12).

Wissensdefizite von Führungskräften und Mitarbeitenden stellen eine weitere Herausforderung bei der Umsetzung dar, weil sie oft nicht über ausreichend Kenntnisse zu strategischen Zielen oder ethischen Anforderungen im Umgang mit digitalen Technologien verfügen vgl. (Dörr, 2020, p. 160). Besonders problematisch ist dies im Umgang mit komplexen digitalen Technologien, deren Wirkungszusammenhänge nicht immer offensichtlich sind: Technologien sind in ihrer Nutzung oft flexibel und dynamisch wandelbar, sodass Unternehmen die langfristigen ethischen Herausforderungen und Auswirkungen, die aus der Nutzung einer Technologie entstehen könnten, nicht vollständig vorhersehen können vgl. (Weißenberger & Marrocco, 2022, p. 50). Zwar können bestimmte Probleme während der Entwicklung und Erfassung von Daten antizipiert werden, doch viele ethische Fragen zeigen sich erst, wenn die Technologie tatsächlich genutzt wird vgl. (Lobschat et al., 2021, p. 876). Auch die Dynamik gesellschaftlicher und politischer Anforderungen, etwa im Bereich Datenschutz, erschwert die Entscheidungsfindung. Nach (Dörr, 2020, p. 165) stellt ein enger kontinuierlicher Austausch mit der Gesellschaft, den Stakeholdern und dem Risikomanagement eine erste Lösung für die Hürden dar.

Ein weiteres Hindernis ist der Einfluss von CDR-Maßnahmen auf den langfristigen Unternehmensgewinn, welcher aufgrund der geringen Evidenz für Manager nicht überzeugend genug ist vgl. (Dörr, 2020, p. 165). Kurzfristige wirtschaftliche Ziele stehen dabei oft im Vordergrund. Dagegen werden CDR-Maßnahmen, die zwar langfristige Vorteile bieten, aber ein schwaches, nicht-finanzielles Ziel darstellen, als weniger dringend angesehen vgl. (Merchant & Van der Stede, 2023, p. 119). Ein solches Phänomen ist in der Fachliteratur auch als Manager-Myopie (Kurzsichtigkeit) bekannt, insbesondere wenn Entscheidungsträger mit Kontrollsystemen incentiviert werden, die einen Fokus auf kurzfristig gewinnbringende Investitionen legen vgl. (Merchant, 1990, p. 308). Dabei ist die Neigung zu einer Kurzsichtigkeit des Managers nicht nur abhängig von seiner Erfahrung oder der verbleibenden Dauer seines Arbeitsverhältnisses vgl. (Narayanan, 1985, p. 1483), sondern auch von seiner persönlichen Haltung gegenüber Maßnahmen, die CSR und damit auch CDR betreffen vgl. (Feder & Weißenberger, 2019, p. 118).

Um diese Barrieren zu überwinden, sind verschiedene Lösungsansätze erforderlich, die sowohl technologische als auch organisatorische Aspekte abdecken. Eine zentrale Maßnahme ist der Aufbau von Wissen und Kompetenzen.

Führungskräfte und Mitarbeitende sollten gezielt zu CDR-relevanten Themen geschult werden, etwa durch interne Trainingsprogramme oder Business-Akademien vgl. (Dörr, 2020, p. 165). Diese Schulungen können dabei helfen, Wissenslücken zu schließen und ein Bewusstsein für die Bedeutung digitaler Verantwortung zu schaffen.

Ein weiterer wichtiger Ansatz ist die Einführung einer klaren Data-Governance. Hierbei geht es um die Festlegung von Richtlinien für den verantwortungsvollen Umgang mit Daten, einschließlich der Ernennung eines Chief-Data-Officers (CDO), der für die Einhaltung dieser Standards verantwortlich ist vgl. (Keimer & Egle, 2020, p. 8). Dieser sollte als Verantwortlicher vor allem sicherstellen, dass die Erhebung und Analyse von Daten den datenschutzrechtlichen Anforderungen und der Compliance entsprechen (Kraft & Drerup, 2020, p. 315). Ergänzend dazu sollten technische Maßnahmen wie Verschlüsselung, Pseudonymisierung und Anonymisierung personenbezogener Daten etabliert werden, um sowohl rechtlichen als auch ethischen Anforderungen gerecht zu werden vgl. (Ruff, 2024, p. 196).

Auch die Wahl geeigneter Technologien zur Datenspeicherung und deren Verwertung spielt eine zentrale Rolle. Im Rahmen der Speicherung ist insbesondere die Entscheidung zwischen *data lakes* und *data warehouses* unabdingbar vgl. (Keimer & Egle, 2020, p. 8). *Data lakes* sind skalierbare Speicherorte für Rohdaten, die sowohl strukturierte als auch unstrukturierte Daten aus verschiedenen Quellen wie mobile Anwendungen oder soziale Medien aufnehmen können.[6] Diese Flexibilität ermöglicht es Unternehmen, Daten in ihrem Rohformat zu speichern und später für spezifische Analysen aufzubereiten. Somit eignen sich *data lakes* besonders für Organisationen, die eine große Menge an heterogenen Datenquellen bewältigen müssen, wie z.B. Technologieunternehmen, die an KI-Lösungen arbeiten, oder *smart cities*, die Daten aus dem Internet der Dinge (IoT) analysieren. Im Gegensatz dazu speichern *data warehouses* Daten in einer festen Struktur oder einem vordefinierten Schema und eignen sich ideal für standardisierte Berichte und historische Analysen. Diese Struktur macht es einfacher, *data warehouses* direkt mit BI-Anwendungen und Dashboards zu verbinden vgl. (Duckstein, 2020, p. 126), was die Nutzung insbesondere für Finanzdienstleister, Handelsunternehmen oder Industrien vorteilhaft macht, die Einblicke aus strukturierten Datenquellen benötigen vgl. (Behme & Mucksch, 2001, p. V).

Die Wahl zwischen diesen Technologien sollte von den spezifischen Anforderungen des Controllings und den geplanten CDR-Maßnahmen abhängen, um eine optimale Unterstützung bei der Datenspeicherung und -analyse zu gewährleisten. Eine Kombination der Vorteile beider Technologien stellt die Verwendung von *lakehouses* dar, die insbesondere im Kontext von *advanced analytics* und Maschinenlernen immer mehr an Bedeutung gewinnen vgl. (Armbrust et al., 2021, p. 1).

[6] Vgl. https://www.intel.de/content/www/de/de/artificial-intelligence/analytics/data-lake-vs-data-warehouse.html, letzter Zugriff am 17.01.2025.

Im Rahmen der Datensammlung und -verwertung ist insbesondere die Nutzung CDR-konformer BI-Software entscheidend, um die Datenqualität sowie den SPOT sicherzustellen sowie Echtzeitanalysen auf Grundlage einer transparenten, verantwortungsvollen Datenbasis zu ermöglichen, die durch ein ERP-System bereitgestellt wird vgl. (Hecht und Scherrer, 2020, p. 86; Keimer und Egle, 2020, p. 9). Diese Technologien unterstützen das Controlling dabei, datenbasierte Entscheidungen schnell und effizient zu treffen, da eine Diskussion über die Wahrheit der Daten erspart bleibt vgl. (Caviezel, 2020, p. 118).

Darüber hinaus sollte das Controlling eine CDR-bewusste Unternehmenskultur fördern, indem sie als integraler Bestandteil der strategischen Unternehmenssteuerung etabliert wird vgl. (Dörr, 2020, p. 196). Ein Smart-Mix aus freiwilligen Selbstverpflichtungen und rechtlichen Vorgaben wie der DSGVO und dem BDSG kann helfen, praktische und ethische Standards zu vereinen (vgl. Lautermann/Frick 2020, S. 51). Transparenz ist dabei essenziell: (Weißenberger & Feder, 2016, p. 18) führen aus, dass im Rahmen eines Management-Approachs die CSR- und damit auch CDR-Aktivitäten strukturiert nach außen kommuniziert werden müssen, um den Eindruck von Greenwashing zu vermeiden und glaubwürdig intrinsische Motivation für nachhaltige Unternehmensziele zu vermitteln. Nach den Autoren bedeutet dies besonders für das Controlling, dass CDR-Maßnahmen bereits in der Planungsphase auf ihre spätere Berichterstattung ausgerichtet sein sollten. Dafür muss CDR als zusätzliche Dimension in die Regelprozesse des Controllings integriert werden, wie etwa in die Zielkataloge, die strategischen Planungsprozesse und die standardisierten Berichtsstrukturen, bspw. durch die Verwendung von Prozess- und Ergebniskennzahlen. (Weißenberger & Feder, 2016, p. 18f.) unterstreichen zudem, dass interaktive Steuerungsprozesse, wie die regelmäßige Überprüfung von CDR-Zielen, eine zentrale Rolle spielen, um Stakeholdern transparente und verlässliche Informationen zu bieten. Auch die Zusammenarbeit mit Abschlussprüfern ist entscheidend, um sicherzustellen, dass CDR-Daten korrekt erfasst werden und ein realistisches Bild der Maßnahmen und Ergebnisse vermittelt wird. Schließlich heben die Autoren hervor, dass die Maßnahmen in leistungsabhängige Entlohnungssysteme und Governance-Strukturen integriert werden sollten, um Transparenz, Nachvollziehbarkeit und die Verankerung von Nachhaltigkeitszielen im Unternehmen zu fördern.

5 Fazit und Ausblick

Diese Arbeit hat gezeigt, dass CDR eine zentrale Rolle für das Controlling spielt und dessen Kernaufgaben auf vielfältige Weise beeinflusst. Im Zuge der digitalen Transformation werden die Domänen des Controllings – Daten, Technologien, Prozesse, Methoden und Kompetenzen – nicht nur verändert, sondern mit neuen Anforderungen konfrontiert. Dadurch wird deutlich, dass CDR als integraler Bestandteil des digitalen Controllings verstanden werden muss, um sowohl

regulatorischen als auch gesellschaftlichen Erwartungen gerecht zu werden.

Der erste Teil der Forschungsfrage, inwiefern CDR die Gestaltung und Umsetzung digitaler Controllingprozesse beeinflusst, konnte somit umfassend beantwortet werden. Es wurde deutlich, dass CDR in jeder der analysierten Domänen eine Rolle spielt, insbesondere in den Bereichen Daten und Kompetenzen. Daten bilden die Grundlage aller Controllingprozesse und erfordern besondere Sorgfalt hinsichtlich des Schutzes, der Qualität und der verantwortungsvollen Nutzung von der Sammlung bis zur Auswertung. Gleichzeitig ist ein nachhaltiges und ethisch orientiertes Handeln nur möglich, wenn die fachlichen und persönlichen Kompetenzen der Controller kontinuierlich weiterentwickelt werden. Ein CDR-orientiertes Mindset im Controlling stellt somit die Basis für nachhaltige und zukunftsorientierte Entscheidungen dar.

Im zweiten Teil der Forschungsfrage, welche Rolle datenschutzrechtliche, ethische und technologische Anforderungen spielen, zeigte sich, dass die Einhaltung von Datenschutzgesetzen wie der DSGVO eine unabdingbare Voraussetzung für die Umsetzung von CDR ist. Darüber hinaus reicht CDR jedoch weit über reine Compliance hinaus: Es verlangt ein menschenzentriertes Denken, das soziale, moralische und ökologische Aspekte einbezieht. Dies wird besonders im Kontext von Big Data und KI deutlich, wo der Umgang mit algorithmischer Verzerrung, Transparenz und ethischen Dilemmas eine der größten Herausforderungen des Jahrzehnts darstellt.

Ein weiterer wichtiger Befund ist, dass die frühzeitige Integration von CDR in das Controlling und die Unternehmensführung nicht nur gesellschaftlichen Anforderungen gerecht wird, sondern auch dazu beiträgt, Datenskandale zu vermeiden und Vertrauen bei Stakeholdern aufzubauen. Dieses Vertrauen kann langfristig einen nachhaltigen Wettbewerbsvorteil schaffen. Gleichzeitig wurde deutlich, dass die Implementierung von CDR mit Herausforderungen verbunden ist. Es bedarf strukturierter Ansätze, wie sie etwa der CDR-Kompass oder das 6-Schritte-Modell nach Dörr bieten, um CDR auf strategischer und operativer Ebene erfolgreich zu implementieren. Die erfolgreiche Umsetzung von CDR erfordert zudem eine proaktive Haltung seitens des Managements und des Controllings. Risiken müssen frühzeitig identifiziert und adressiert werden, wobei Scheitern und kontinuierliches Lernen Teil des Prozesses sind. Für das Controlling bedeutet dies, dass es über klassische Aufgaben hinausgeht und eine gestaltende sowie überwachende Rolle einnimmt. Controller werden zu Akteuren, die nicht nur Daten analysieren, sondern auch ethische Prinzipien und Nachhaltigkeitsaspekte in ihre Arbeit integrieren. Die Einführung klarer Prozesse zur Datennutzung und die Förderung eines ethischen Leitbildes tragen dazu bei, die Unternehmenssteuerung im Zeitalter von KI und Big Data nachhaltig auszurichten.

Obwohl diese Arbeit die Bedeutung von CDR für das digitale Controlling umfassend analysiert hat, bestehen auch einige Grenzen. Die theoretische Grundlage zur CDR-Implementierung im Controlling befindet sich noch in einem frühen Stadium, wodurch empirische Daten und Fallstudien nur begrenzt zur Verfügung standen. Dies schränkt die Möglichkeit ein, praktische Handlungsempfehlungen vollständig zu validieren. Zudem bleibt offen, in welchem Umfang Unternehmen bereit sind, die notwendigen Investitionen in grüne IT-Infrastrukturen und Mitarbeiterschulungen zu tätigen, um CDR erfolgreich umzusetzen. Ein weiterer kritischer Punkt ist die Definition von CDR als freiwillige Verantwortung, wie sie etwa vom BMUV verwendet wird.[7] Diese Freiwilligkeit wird zunehmend durch regulatorische Vorgaben wie die DSGVO eingeschränkt, was die Frage aufwirft, ob CDR langfristig ein Alleinstellungsmerkmal für Unternehmen sein kann oder lediglich eine Erweiterung bestehender Compliance-Anforderungen darstellt vgl. (Ruff, 2024, p. 206).

Weiterhin sollte untersucht werden, wie sich CDR langfristig auf die Reputation und den wirtschaftlichen Erfolg von Unternehmen auswirkt. Empirische Studien könnten dabei helfen, die Frage zu klären, ob die finanziellen und gesellschaftlichen Vorteile einer CDR-Implementierung die hohen Anfangsinvestitionen rechtfertigen. Zudem wäre es wertvoll, konkrete Metriken zu entwickeln, um den Erfolg von CDR-Maßnahmen im Controlling messbar zu machen. Darüber hinaus könnte untersucht werden, wie regulatorische Anforderungen die freiwillige Umsetzung von CDR beeinflussen und ob diese Einschränkungen die Akzeptanz bei Unternehmen fördern oder hemmen. Abschließend bleibt zu klären, wie KMU mit begrenzten Ressourcen CDR effektiv umsetzen können. Insgesamt bietet das Thema CDR im digitalen Controlling ein breites Feld für weitere Forschung und praktische Anwendung, das in den kommenden Jahren zunehmend an Bedeutung gewinnen wird.

Literatur

Agudelo, M. A. L., Jóhannsdóttir, L., & Davídsdóttir, B. (2019). A Literature Review Of The History And Evolution Of Corporate Social Responsibility. *International Journal of Corporate Social Responsibility*, *4*(1), 1–23.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (Xai): What We Know And What Is Left To Attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, 1–52.

Andersen, N. (2022). Corporate Digital Responsibility Survey 2022 [Accessed: 13.10.2024]. https://www.deloitte.com

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A New Generation Of Open Platforms That Unify Data Warehousing And Advanced Analytics. *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR '21)*.

Behme, W., & Mucksch, H. (2001). *Data Warehouse-gestützte Anwendungen. Theorie und Praxiserfahrungen in verschiedenen Branchen* (1. Aufl.). Springer Vieweg.

Bharadiya, J. P. (2023). The Role Of Machine Learning In Transforming Business Intelligence. *International Journal of Computing and Artificial Intelligence*, *4*(1), 16–24.

Binder, U. (2017). Controlling Aufgaben, Funktionen Und Ziele [Accessed: 30.11.2024]. https://www.haufe.de/controlling/controllerpraxis/was-ist-controlling/controlling-aufgaben-funktionen-ziele_112_433744.html

[7] Vgl. https://cdr-initiative.de/handlungsfelder/noch-ein-overlayer, letzter Zugriff am 20.01.2025.

Bowen, H. R. (1953). *Social Responsibilities Of The Businessman*. Harper & Brothers.

Carroll, A. B. (1991). The Pyramid Of Corporate Social Responsibility: Toward The Moral Management Of Organizational Stakeholders. *Business Horizons*, *34*(4), 39–48.

Carroll, A. B. (1999). Corporate Social Responsibility. *Business & Society*, *38*(3), 268–295.

Caviezel, R. (2020). Mit Business Intelligence Die Unternehmenssteuerung Digitalisieren. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 103–121). Springer Gabler.

Corporate Digital Responsibility-Initiative. (2021). Corporate Digital Responsibility-Kodex [Accessed: 26.11.2024]. https://cdr-initiative.de/uploads/files/2024-01_Kodex_CDR-Initiative.pdf

Datenschutz-Grundverordnung (DSGVO). (2016). General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 [Protection of natural persons with regard to the processing of personal data]. https://eur-lex.europa.eu

Davenport, T. H., & Bean, R. (2017). *Big Data Executive Survey 2017. Big Data Business Impact: Achieving Business Results through Innovation and Disruption* (Techn. Ber.). NewVantage Partners.

Dörr, S. (2020). *Praxisleitfaden Corporate Digital Responsibility: Unternehmerische Verantwortung Und Nachhaltigkeitsmanagement Im Digitalzeitalter* (1. Aufl.). Springer Gabler.

Duckstein, R. (2020). Von Digitalen Hilfsmitteln Zur Digitalen Methodik. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 123–139). Springer Gabler.

Engelbers, J., & Moreira, D. (2020). Wie Zalando digitale Lösungen nutzt, um das Investment-Controlling zu transformieren. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 287–302). Springer Gabler.

Feder, M., & Weißenberger, B. E. (2019). Understanding the behavioral gap: Why would managers (not) engage in CSR-related activities? *Journal of Management Control*, *30*, 95–126.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). Ai4People—An Ethical Framework For A Good Ai Society: Opportunities, Risks, Principles, And Recommendations. *Minds and Machines*, *28*, 689–707.

Gerig, I. (2020). Standardisierung und Automatisierung als Basis für die Digitalisierung im Controlling von Siemens Building Technologies. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 211–236). Springer Gabler.

Gotsch, M., Martin, N., Eberling, E., Shirinzadeh, S., Kuhlmann, D., Petschow, U., & Pentzien, J. (2022). *Der Beitrag Von Big Data, Ki Und Digitalen Plattformen Auf Dem Weg Zu Einer Green Economy - Einsatzbereiche Und Transformationspotenziale* (Techn. Ber.). Umweltbundesamt.

Hecht, N., & Scherrer, P. (2020). Nutzen und Stolpersteine bei der Einführung einer Business Intelligence-Lösung für KMU am Beispiel der Firma SIGA. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 83–102). Springer Gabler.

Keimer, I., & Egle, U. (Hrsg.). (2020). *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (1. Aufl.). Springer Gabler.

Kraft, M., & Drerup, B. (2020). Digitalisierung Des Controllings In Versicherungsunternehmen. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 303–321). Springer Gabler.

Kunz, W. H., & Wirtz, J. (2024). Corporate Digital Responsibility (Cdr) In The Age Of Ai: Implications For Interactive Marketing. *Journal of Research in Interactive Marketing*, *18*(1), 31–37.

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate Digital Responsibility. *Journal of Business Research*, *122*, 875–888.

Manhart, K. (2024). Datenverarbeitung: Wie Big Data Nicht Zu Teuer Wird. Verfügbar 2. Januar 2025 unter https://www.golem.de/news/datenverarbeitung-wie-big-data-nicht-zu-teuer-wird-2412-187845.html

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, *90*(10), 1–9.

Merbecks, U. (2023). Corporate Digital Responsibility In Germany [Accessed: 11.10.2024]. *Beiträge des Forschungs- und Transfertags der Hochschule Rhein-Waal*. https://opus4.kobv.de/opus4-rhein-waal/frontdoor/index/index/docId/1763

Merchant, K. A. (1990). The Effects Of Financial Controls On Data Manipulation And Management Myopia. *Accounting, Organizations and Society*, *15*(4), 297–313.

Merchant, K. A., & Van der Stede, W. (2023). *Management Control Systems* (5. Aufl.). Pearson.

Möller, K., Seefried, J., & Wirnsperger, F. (2017). Wie Controller Zu Business-Partnern Werden. *Controlling & Management Review*, *61*, 64–67.

Narayanan, M. P. (1985). Managerial Incentives For Short-Term Results. *The Journal of Finance*, *40*(5), 1469–1484.

Noti, K., Mucciarelli, F. M., Angelici, C., dalla Pozza, V., & Pillinini, M. (2020). *Corporate Social Responsibility (Csr) And Its Implementation Into Eu Company Law* (Techn. Ber.) (Accessed: 20.10.2024). European Parliament, Policy Department for Citizens' Rights und Constitutional Affairs. https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)658541

OHCHR. (2011). *Guiding Principles On Business And Human Rights* (Techn. Ber.). Office of the High Commissioner for Human Rights.

Pick, D., & Knopf, T. (2023). Corporate Digital Responsibility - Begriff, Motive und Maßnahmen von Handelsunternehmen. In M. Bruhn & K. Hadwig (Hrsg.), *Gestaltung Des Wandels Im Dienstleistungsmanagement, Forum Dienstleistungsmanagement* (S. 672–697). Springer Gabler.

Rai, A. (2020). Explainable Ai: From Black Box To Glass Box. *Journal of the Academy of Marketing Science*, *48*, 137–141.

Rödder, N. (2018). Transparenz Im Datenschutz. In Smart-Data-Begleitforschung (Hrsg.), *Corporate Social Responsibility – Fachgruppe Wirtschaftliche Potenziale & gesellschaftliche Akzeptanz* (S. 14–15). Berlin.

Ruff, D. (2024). Corporate Digital Responsibility Im Kontext Eines Entstehenden Datenrechts. In S. Laimer, A.-C. Mittwoch, T. Müller & L. Staffler (Hrsg.), *Daten, Plattformen, Smart Contracts. Aktuelle Rechtsfragen* (1. Aufl., S. 167–213). Nomos.

Schäffer, U., & Weber, J. (2016a). Die Digitalisierung wird das Controlling radikal verändern. *Controlling & Management Review*, *60*, 6–17.

Schäffer, U., & Weber, J. (2016b). *Einführung in das Controlling* (15. Aufl.). Schäffer-Poeschel.

Thorun, C. (2018). Corporate Digital Responsibility: Unternehmerische Verantwortung In Der Digitalen Welt. In C. Gärtner & C. Heinrich (Hrsg.), *Fallstudien zur Digitalen Transformation: Case Studies für die Lehre und praktische Anwendung* (S. 173–191). Springer Gabler.

Trachsel, V., & Bitterli, C. (2020). Controller-Profile In Der Schweiz - Bedeutung Der Digitalisierung. In I. Keimer & U. Egle (Hrsg.), *Die Digitalisierung Der Controlling-Funktion: Anwendungsbeispiele Aus Theorie Und Praxis* (S. 199–210). Springer Gabler.

Van der Merwe, J., & Al Achkar, Z. (2022). Data Responsibility, Corporate Social Responsibility, And Corporate Digital Responsibility. *Data & Policy*, *4*, e12.

Weißenberger, B. E., & Bauch, K. A. (2019). Controlling In Der Echtzeit-Economy: Auswirkungen Der Digitalen Transformation Auf Die Unternehmenssteuerung. In R. Obermaier (Hrsg.), *Handbuch Industrie 4.0 und Digitale Transformation: Betriebswirtschaftliche, technische und rechtliche Herausforderungen* (S. 703–721). Springer Gabler.

Weißenberger, B. E., & Feder, M. (2016). CSR richtig managen: Außenkommunikation und Controlling gehören dazu. *Personal Quarterly*, *68*(4), 15–19.

Weißenberger, B. E., & Marrocco, A. (2022). Corporate Digital Responsibility und ihre Integration in die Unternehmensführung. In H. Corsten & S. Roth (Hrsg.), *Handbuch Digitalisierung* (S. 41–58). Vahlen.

Willrich, S. (2018). Abgrenzung Zur Corporate Social Responsibility. In Smart-Data-Begleitforschung (Hrsg.), *Corporate Social Responsi-*

bility – Fachgruppe Wirtschaftliche Potenziale & gesellschaftliche Akzeptanz (S. 8). Berlin.

Abkürzungsverzeichnis

Abs. Absatz

Art. Artikel

BA Business Analytics

BDSG Bundesdatenschutzgesetz

BI Business Intelligence

BMUV Bundesministerium für Umwelt- und Verbraucherschutz

Buchst. Buchstabe

CDO Chief Data Officer

CDR Corporate Digital Responsibility

CSR Corporate Social Responsibility

DAX Deutscher Aktienindex

DSGVO Datenschutzgrundverordnung

ERP Enterprise Resource Planning

IoT Internet of Things

IT Informationstechnologie

KI Künstliche Intelligenz

KMU Kleine/Mittlere Unternehmen

KPI Key Performance Indicator

OHCHR Office of the High Commissioner for Human Rights (Büro des Hohen Kommissars für Menschenrechte)

SPOT Single Point of Truth

UN United Nations (Vereinte Nationen)

XAI Explainable Artificial Intelligence

Junior Management Science 11(1) (2026) 43-73

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Textauswertung in der Finanzanalyse: Eine Ereignisstudie zum Sentiment in Earnings Calls

Leonie Terhardt

Heinrich Heine University Düsseldorf

Abstract

This thesis examines the influence of sentiment in earnings calls on financial market reactions. The objective is to analyse the extent to which the tone used in presentations and Q&A segments affect abnormal returns, and how differences in investors' information environments shape these effects. The study is based on transcripts from companies in the S&P Composite 1500, whose tone is recorded using finance-specific lexicons and evaluated within an event-study framework. The results suggest that sentiment in the Q&A section in particular triggers significant immediate market reactions, with analysts' questions having the greatest influence. Conversely, the sentiment of the presentation assumes significance during the drift period and frequently results in subsequent corrections. In addition, investors show increased sensitivity to verbal signals when it comes to companies with a higher degree of uncertainty. The thesis contributes to the literature by demonstrating that context-specific sentiment measurement provides valuable insights for the evaluation of qualitative company information. Furthermore, it highlights the limitations of generic text lexicons.

Zusammenfassung

Die vorliegende Arbeit untersucht den Einfluss des Sentiments in Earnings Calls auf die Reaktionen der Finanzmärkte. Ziel ist es zu analysieren, inwiefern der sprachliche Ausdruck in der Präsentation und im Rahmen des Q&A-Segments abnormale Renditen beeinflusst und welche Bedeutung die Informationslage der Investoren dabei hat. Die Untersuchung basiert auf Transkripten von Unternehmen des S&P Composite 1500, deren Tonalität mithilfe finanzspezifischer Lexika erfasst und im Rahmen einer Ereignisstudie ausgewertet wird. Die Ergebnisse der Untersuchung legen nahe, dass insbesondere das Sentiment im Q&A-Teil signifikante unmittelbare Marktreaktionen auslöst, wobei die Fragen der Analysten den größten Einfluss besitzen. Demgegenüber gewinnt das Sentiment der Präsentation in der Driftperiode an Bedeutung und führt häufig zu späteren Korrekturen. Darüber hinaus zeigen Investoren in Bezug auf Unternehmen mit einer höheren Unsicherheit eine erhöhte Sensibilität auf verbale Signale. Die Arbeit leistet einen Beitrag zur Forschung, indem sie zeigt, dass eine kontextspezifische Sentimentmessung zentrale Einblicke für die Bewertung qualitativer Unternehmensinformationen liefert und dabei die Grenzen generischer Textlexika aufzeigt.

Keywords: sentiment; earnings calls; textanalyse; ereignisstudie; finanzmarktreaktionen

1 Einleitung

Während traditionell finanzielle Kennzahlen wie die Gewinne pro Aktie im Fokus standen, wird in der Finanzanalyse zunehmend erkannt, dass diese allein nicht ausreichen, um ein umfassendes Bild der Unternehmensleistung zu erhalten.[1] Für die Einschätzung von Chancen, Risiken sowie der strategischen Unternehmensausrichtung sind vor allem qualitative Informationen aus Unternehmensveröffentlichungen von Bedeutung. Vor diesem Hintergrund wächst das Interesse an der systematischen Auswertung sprachlicher Inhal-

[1] Vgl. Davis und Tama-Sweet (2012, p. 806); Feldman et al. (2009, p. 916).

DOI: https://doi.org/10.5282/jums/v11i1pp43-73

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

te der Unternehmenskommunikation. Ein besonders vielversprechender Ansatz in diesem Bereich ist die Sentimentanalyse, die es ermöglicht, Stimmungen in Texten zu quantifizieren.[2] Insbesondere Earnings Calls – Telefonkonferenzen, in denen Unternehmen ihre finanziellen Ergebnisse präsentieren und sich den Fragen von Analysten stellen – erweisen sich hierbei als wertvolle Informationsquelle.[3] Die dynamische Interaktion, insbesondere im Frage- und Antwortabschnitt, liefert oft wertvolle qualitative Informationen, die über die rein formalen Berichterstattungen hinausgehen und subtile Hinweise auf die Stimmung der Unternehmensvertreter und Analysten sowie auf die Zukunftsaussichten des Unternehmens geben.

Die vorliegende Arbeit zielt darauf ab, aufzuzeigen, wie die Textauswertung als ergänzendes Instrument zur klassischen Finanzanalyse eingesetzt werden kann. Konkret wird untersucht, welche Rolle die Sentimentanalyse im Kontext von Earnings Calls spielt, um ein tieferes Verständnis für den Zusammenhang zwischen den in Unternehmensmitteilungen ausgedrückten Stimmungen und dem Marktverhalten zu erlangen. Angesichts widersprüchlicher Forschungsergebnisse soll insbesondere geklärt werden, wie das Sentiment in verschiedenen Abschnitten von Earnings Calls die Marktreaktionen beeinflusst und welche Rolle dabei die Informationslage der Anleger sowie methodische Unterschiede in der lexikonbasierten Sentimentanalyse spielen.

Hierzu ist die Arbeit folgendermaßen aufgebaut: Im zweiten Kapitel werden die theoretischen Grundlagen und der aktuelle Forschungsstand der textbasierten Sentimentanalyse sowie deren Anwendung auf Earnings Calls dargelegt. Dabei wird die Bedeutung qualitativer Inhalte in Earnings Calls als wesentliche Informationsquelle erörtert und aufgezeigt, wie das in Texten enthaltene Sentiment ausgewertet werden kann. Das dritte Kapitel beschreibt die methodische Vorgehensweise. Anhand einer Stichprobe von Earnings Calls aus dem S&P Composite 1500 werden mit einem lexikonbasierten Ansatz Sentiment-Indikatoren für verschiedene Abschnitte und Redner ermittelt. Um die Marktreaktionen auf das Sentiment der Earnings Calls zu analysieren, werden darüber hinaus mittels einer Ereignisstudien-Methodik abnormale Renditen berechnet. Im vierten Kapitel werden die Ergebnisse der Auswirkungen des Sentiments in verschiedenen Segmenten von Earnings Calls auf die Marktreaktionen dargestellt und diskutiert. Ein besonderer Fokus liegt hierbei auf den Unterschieden, die sich je nach verwendetem Lexikon zur Ermittlung ergeben, um die Sensitivität gegenüber unterschiedlichen textanalytischen Methoden zu beleuchten. Darüber hinaus wird der Einfluss der Informationslage auf die Reaktionen der Investoren und die Robustheit der Ergebnisse überprüft. Die Arbeit schließt im fünften Kapitel mit einer Zusammenfassung der wichtigsten Erkenntnisse sowie Empfehlungen für zukünftige Forschungsarbeiten ab.

Die Ergebnisse der Arbeit bestätigen, dass emotionale und sprachliche Nuancen aus Earnings Calls eine zentrale Rolle spielen und mithilfe der Sentimentanalyse systematisch erfasst werden können. Hierbei hat der Q&A-Abschnitt, insbesondere die Fragen der Analysten, einen besonders starken Einfluss auf die Marktreaktionen. Nach einer anfänglich emotional geprägten Reaktion beginnen Investoren, die langfristigen Implikationen der erhaltenen Informationen genauer zu analysieren, was in der Driftperiode häufig zu einer negativen Korrektur der anfänglich positiven Marktreaktionen führt. Die Reaktionen der Anleger hängen dabei stark von ihrer Informationslage ab, wobei Unternehmen mit höherer Unsicherheit stärkere Marktreaktionen auf das Sentiment zeigen. Erfahrene Investoren sind besser in der Lage, wertrelevante Informationen aus dem Ton von Earnings Calls zu extrahieren, während unerfahrene Anleger ihre initiale Reaktion häufiger revidieren. Der Einfluss der einzelnen Abschnitte und Redner eines Earnings Calls variiert ebenfalls je nach Informationslage. Insbesondere die Antworten des Managements im Q&A-Abschnitt gewinnen bei zunehmenden Informationsasymmetrien an Bedeutung. Auch erfahrene Anleger messen ihnen eine größere Relevanz bei. Die Arbeit unterstreicht zudem die methodischen Herausforderungen einer lexikonbasierten Sentimentanalyse bei der Erfassung von Kontext und Mehrdeutigkeit. Hierbei wird insbesondere deutlich, dass für eine präzise Auswertung der Texte jedes Segment eines Earnings Calls ein spezifisch angepasstes Lexikon erfordert, und nicht auf ein universelles Lexikon zurückgegriffen werden sollte.

2 Theoretische Grundlagen der Sentimentanalyse in Earnings Calls

2.1 Earnings Calls als Medium der qualitativen Offenlegung

Im zweiten Quartal 2019 veröffentlichte Netflix Inc. starke Umsatz- und Gewinnzahlen, die auf den ersten Blick eine positive Entwicklung des Unternehmens signalisierten.[4] Während des Earnings Calls gab das Unternehmen jedoch bekannt, dass die Anzahl neuer Abonnenten hinter den Erwartungen zurückblieb und äußerte Bedenken hinsichtlich des zukünftigen Wachstums durch zunehmenden Wettbewerbsdruck. Diese qualitativen Informationen führten trotz positiver Finanzkennzahlen zu einem Kursverlust von über 10 %.

Das angeführte Beispiel verdeutlicht die zentrale Bedeutung qualitativer Informationen in der Finanzkommunikation und deren potenziellen Einfluss auf die Marktdynamik, unabhängig davon, ob die quantitativen Ergebnisse positiv oder negativ ausfallen. Die finanzielle Berichterstattung eines Unternehmens zielt darauf ab, Informations-asymmetrien und Interessenkonflikte zwischen Managern und externen Stakeholdern zu reduzieren.[5] Sie umfasst quantitative Kennzahlen und qualitative Einschätzungen, sowie sowohl verpflichtende, als auch freiwillige

[2] Vgl. Todd et al. (2024, p. 2).
[3] Vgl. De Amicis et al. (2021); Blau et al. (2015, p. 204); Brockman et al. (2015, pp. 24f).
[4] Vgl. Rodriguez (2019).
[5] Vgl. Davis und Tama-Sweet (2012, p. 804).

Offenlegungen. In der Vergangenheit hat sich Finanzanalyse primär auf numerische Aspekte konzentriert.[6] Dies ist darauf zurückzuführen, dass quantitative Daten im Vergleich zu qualitativen leichter verfügbar, objektiver und eindeutiger sind. Zahlen allein können jedoch das gesamte Spektrum der Unternehmensleistung nicht vollständig abbilden.[7] Qualitative Informationen, insbesondere die Stimmungen und Einschätzungen des Managements, spielen eine entscheidende Rolle bei der Interpretation von Chancen, Risiken und der allgemeinen Unternehmensstrategie. Die Integration von Aspekten der qualitativen Finanzberichterstattung in die Finanzanalyse eröffnet daher die Möglichkeit, ein ganzheitlicheres Bild der Finanzmarktdynamik zu zeichnen und hilft Investoren fundiertere Entscheidungen zu treffen.[8]

Die zunehmende Bedeutung qualitativer Informationen in der Finanzberichterstattung wird insbesondere anhand der Entwicklung sogenannter Earnings Calls deutlich. Hierbei handelt es sich um eine Präsentation und Diskussion der Finanzergebnisse eines Unternehmens durch dessen Management in Form einer Telefonkonferenz.[9] Diese Konferenzen finden auf freiwilliger Basis kurz nach der Veröffentlichung der jeweiligen Quartals- oder Jahresergebnisse statt. Das Inkrafttreten der Regulation Fair Disclosure (Reg FD) im Jahr 2000 hat zu einer deutlich stärkeren Nutzung dieser Plattform durch Unternehmen für die Kommunikation mit Investoren geführt und den Informationsgehalt der Telefonkonferenzen erhöht.[10] Die Reg FD verpflichtet börsennotierte Unternehmen dazu, wichtige, bislang nicht öffentlich bekannte, Informationen zu veröffentlichen. Vor dieser Regulierung waren Earnings Calls nur für ausgewählte große Investoren und Analysten über private Telefonnummern zugänglich.[11] Seit Einführung der Reg FD muss die Teilnahme auch der breiten Öffentlichkeit ermöglicht werden. Gegenwärtig erfolgt die Übertragung von Earnings Calls vielfach live über Investor-Relations-Websites oder Streaming-Dienste wie YouTube. Die simultane Übermittlung von Informationen gewährleistet eine gerechte Informationsverteilung und löst damit das Problem selektiver Offenlegungen.[12]

Im Rahmen der initialen Forschungsarbeiten zum Stellenwert von Earnings Calls erfolgte zunächst eine Untersuchung der grundlegenden Nutzung dieser Kommunikationsform, da seinerzeit nur eine unzureichende Datenbasis zu den tatsächlich kommunizierten Inhalten verfügbar war.[13] Frankel et al. (1999) identifizierten Telefonkonferenzen als einflussreiches Informationsereignis, das durch signifikant hohe Renditevolatilität und Handelsvolumen gekennzeichnet ist.[14] In einer weiteren Untersuchung zeigten Bowen et al. (2002), dass die Prognosegenauigkeit von Analysten nach Gewinnankündigungen mit Earnings Calls im Vergleich zu solchen ohne Earnings Call zunimmt.[15] Kimbrough (2005) konnte darüber hinaus nachweisen, dass der Markt auf aktuelle Gewinne weniger sensibel reagiert, wenn eine Gewinnankündigung mit einer Telefonkonferenz einhergeht, als wenn dies nicht der Fall ist.[16]

Ein Earnings Call beginnt typischerweise mit einer sogenannten „Safe Harbour"-Erklärung des Managements, in der darauf hingewiesen wird, dass die präsentierten Daten zukunftsgerichtete Aussagen enthalten, die von den tatsächlichen Ergebnissen abweichen können.[17] Earnings Calls umfassen eine Präsentation sowie einen Frage-und-Antwort-Abschnitt (Q&A).[18] Im Rahmen der Präsentation, die in der Regel eine Dauer von 15 bis 20 Minuten hat, stellen die Unternehmensvertreter die Ergebnisse des abgelaufenen Berichtszeitraums dar. In der Regel übernimmt der Chief Executive Officer (CEO) die Berichterstattung über die allgemeine Geschäfts- und Marktperformance sowie die strategische Vision, während der Chief Financial Officer (CFO) für die Erläuterung der finanziellen Details und Hintergründe zuständig ist. Weitere Führungskräfte können spezifische Einblicke in ihre Verantwortungsbereiche geben. Das anschließende Q&A erstreckt sich über einen Zeitraum von 30 bis 45 Minuten und bietet Analysten und anderen Teilnehmern die Möglichkeit, die präsentierten Informationen zu hinterfragen oder zusätzliche Details zu verlangen, die nicht in der Präsentation oder in anderen Offenlegungskanälen wie Geschäftsberichten oder Pressemitteilungen deutlich geworden sind. Die Unternehmensleitung kann durch gezielte Anfragen seitens der Teilnehmer zur Preisgabe privater Informationen veranlasst werden, einschließlich Angaben, die von den Managern lieber zurückbehalten werden würden. Manager sind zwar nicht verpflichtet, auf direkte Anfragen von Analysten zu antworten, jedoch besteht die Möglichkeit, dass der Markt aus einer Weigerung, Informationen zu veröffentlichen, negative Schlussfolgerungen zieht.[19]

Im Gegensatz zu standardisierten und vorbereiteten Dokumenten wie Finanzberichten oder Pressemitteilungen ist das Informationsumfeld einer Telefonkonferenz demnach dynamischer und unstrukturierter.[20] Insbesondere der spontane Charakter des Q&A-Abschnitts stellt eine Herausforderung für die Manager dar, da sie gleichzeitig den Inhalt und die Sprache ihrer Aussagen kontrollieren müssen.[21] Über die inhaltlich neuen Offenlegungen hinaus bieten die qualitativen Aspekte dieser Gespräche potenziell informative verbale Hinweise, die auch emotionale Ausdrücke wie „zufrieden" oder „ernüchtert" umfassen können.[22] Die Wortwahl

[6] Vgl. Feldman et al. (2009, p. 916).
[7] Vgl. Davis und Tama-Sweet (2012, p. 806).
[8] Vgl. Feldman et al. (2009, p. 916).
[9] Vgl. CFI Education Inc. (2024).
[10] Vgl. Brockman et al. (2015, p. 24); Irani (2004, p. 26).
[11] Vgl. Elliott et al. (2023, p. 11).
[12] Vgl. Hollander et al. (2010, p. 536); Frankel et al. (1999, p. 135).
[13] Vgl. Matsumoto et al. (2011, p. 1386).
[14] Vgl. Frankel et al. (1999, pp. 134f).
[15] Vgl. Bowen et al. (2002, pp. 285f).
[16] Vgl. Matsumoto et al. (2011, p. 1386).
[17] Vgl. CFI Education Inc. (2024).
[18] Vgl. Hollander et al. (2010, p. 537).
[19] Vgl. Hollander et al. (2010, p. 559).
[20] Vgl. Blau et al. (2015, p. 206).
[21] Vgl. De Amicis et al. (2021).
[22] Vgl. Elliott et al. (2023, p. 8); Matsumoto et al. (2011, p. 1384).

der Teilnehmer kann stark nuanciert sein, da Manager versuchen, das Unternehmen positiv darzustellen, ohne negative Konsequenzen zu riskieren.[23] Im Vergleich zu Zahlen bieten Worte Managern bei qualitativen Offenlegungen dabei einen größeren Spielraum. Obgleich sie sich der Problematik bewusst sind, offensichtlich falsche Aussagen zu treffen, ist die Formulierung von Faktenbehauptungen in positiveren oder weniger positiven Ausdrücken mit einem geringeren Risiko für regulatorische Interventionen verbunden.[24] Earnings Calls werden von Managern als weniger formell und haftungsintensiv wahrgenommen als Pressemitteilungen.[25] Infolgedessen sind sie während dieser Telefonkonferenzen eher geneigt, bestimmte Informationen preiszugeben, die sie in einer Pressemitteilung zurückhalten würden. Insbesondere dann, wenn spezifische Details in den formelleren Berichterstattungen nicht offengelegt werden können, kann der Tonfall in diesen Konferenzen eine Signalwirkung entfalten.[26] Earnings Calls zeichnen sich demnach durch eine subtile Informationsvermittlung aus, die durch traditionelle Methoden der Finanzanalyse nur schwer zu erfassen und auszuwerten ist.[27]

2.2 Sentimentanalyse im Kontext von Markteffizienz und Behavioural Finance

Der Informationsgehalt von Earnings Calls wirft Fragen bezüglich der Markteffizienz auf, wie sie in der Effizienzmarkthypothese (EMH) beschrieben wird. Gemäß der EMH reflektieren die Preise von Vermögenswerten ihren fundamentalen Wert und enthalten sämtliche relevante Informationen, die zu einem gegebenen Zeitpunkt von rationalen Marktakteuren effizient und fehlerfrei genutzt werden.[28] Die EMH wird in drei Formen unterteilt: die starke, die semi-starke und die schwache Form. Die starke Form besagt, dass alle Informationen, sowohl öffentliche als auch private, in den Preisen enthalten sind. Dies impliziert, dass selbst Insider keinen Vorteil haben und keine überdurchschnittlichen Renditen erzielen können, da alle Informationen bereits vollständig im Preis integriert sind. Demzufolge sollte die Veröffentlichung von Earnings Calls keinen signifikanten Einfluss auf die Aktienkurse haben. Treten dennoch um Earnings Calls herum außergewöhnliche Renditen auf, lässt dies den Schluss zu, dass die Märkte nicht vollständig effizient sind und Informationen nicht unverzüglich oder vollständig eingepreist wurden. Die semi-starke Form der EMH modifiziert die ursprüngliche Hypothese, indem sie annimmt, dass sämtliche öffentlich verfügbaren Informationen – darunter vergangene Preisbewegungen, Unternehmensberichte und Nachrichten – im aktuellen Preisniveau abgebildet sind.[29] Es wird also angenommen, dass Anleger auf neu veröffentlichte, unternehmensspezifische Informationen reagieren.[30] Dies eröffnet die Möglichkeit, dass sie ihre Erwartungen und Anlageentscheidungen basierend auf den Signalen in Earnings Calls anpassen, sofern diese über bereits bekannte oder antizipierte Informationen hinausgehen.[31] In ihrer schwachen Form besagt die EMH demgegenüber lediglich, dass alle historischen Informationen in den aktuellen Preisen reflektiert sind.[32]

Im Gegensatz zur EMH stellt die Behavioral Finance die Annahme der vollkommen rationalen Marktakteure infrage und integriert psychologische Erkenntnisse in die Untersuchung von Finanzmärkten.[33] Sie nimmt an, dass die Preisbildungsmechanismen nicht nur von historischen, öffentlichen und privaten Informationen, sondern auch stark von individuellen Ansichten und Emotionen der Anleger beeinflusst werden. Soziale, verhaltensbedingte und psychologische Faktoren können die Erwartungen und Entscheidungen von Menschen maßgeblich beeinflussen, sodass allgemein bekannte Informationen nicht rational verarbeitet werden. Emotionen und kognitive Verzerrungen können demnach zu einer temporären Fehlbewertung des Werts von Vermögenswerten führen,[34] wobei Unsicherheit als verstärkender Faktor wirkt.[35] Die von der Behavioral Finance beschriebenen Verzerrungen im Anlegerverhalten stehen im Widerspruch zur EMH, gemäß derer Abweichungen zwischen Marktpreisen und ihren fundamentalen Werten durch rationale Marktteilnehmer, sogenannte Arbitrageure, zeitnah korrigiert und irrationale Händler vom Markt verdrängt werden.[36] Es konnte jedoch festgestellt werden, dass rationale Händler in der Regel risikoscheu sind, über einen kurzen Handelshorizont verfügen und geringere Renditen erzielen.[37] In der Konsequenz des begrenzten Handlungsspielraums rational handelnder Investoren kann es somit durch das Verhalten irrationeller Investoren zu Marktineffizienzen und fehlerhaften Preisbildungen am Kapitalmarkt kommen. [38]

Vor diesem Hintergrund nimmt die Untersuchung der Auswirkung von Sentiment auf finanzielle Entscheidungen in der wissenschaftlichen Forschung zunehmend an Bedeutung zu.[39] Der Begriff „Sentiment" wird im allgemeinen Sprachgebrauch verwendet, um Gefühle und Überzeugungen gegenüber einem bestimmten Sachverhalt oder einer Situation zu beschreiben[40] und wird in der deutschsprachigen Literatur häufig mit „Stimmung" übersetzt.[41] Die Sentimentanalyse zielt darauf ab, Meinungen, Einstellungen

[23] Vgl. Blau et al. (2015, p. 204).
[24] Vgl. Davis und Tama-Sweet (2012, p. 644).
[25] Vgl. Matsumoto et al. (2011, p. 1387)
[26] Vgl. Pouryousof et al. (2023).
[27] Vgl. Blau et al. (2015, p. 204).
[28] Vgl. Malkiel und Fama (1970, p. 383).
[29] Vgl. Malkiel und Fama (1970, p. 383).
[30] Vgl. Blau et al. (2015, p. 203).
[31] Vgl. Rahman (2023).
[32] Vgl. Malkiel und Fama (1970, p. 383).
[33] Vgl. Baddeley (2018, pp. 9, 14).
[34] Vgl. H. K. Baker und Nofsinger (2010, p. 340).
[35] Vgl. Baddeley (2018, p. 5).
[36] Vgl. H. K. Baker und Nofsinger (2010, p. 595).
[37] Vgl. Gregory (2021).
[38] Vgl. M. Baker und Wurgler (2007, p. 131).
[39] Vgl. M. Baker und Wurgler (2007, p. 131).
[40] Vgl. Liu (2020, p. 2).
[41] Vgl. boerse.de (n. d.).

und Emotionen zu identifizieren und zu quantifizieren.[42] In finanziellen Analysen fungiert sie als Bindeglied zwischen der Effizienzmarkthypothese und der Behavioral Finance, indem sie die Stimmungslage der Marktteilnehmer untersucht. Auf diese Weise lässt sich aufzeigen, wie emotionale und kognitive Verzerrungen die Märkte beeinflussen und zu Abweichungen von der rationalen Marktlogik der Effizienzmarkthypothese führen. Die gegenwärtige wissenschaftliche Literatur beschäftigt sich daher intensiv mit der Messung und Analyse des Einflusses von Sentiment auf das Verhalten von Marktteilnehmern sowie dessen Bedeutung für die Prognosefähigkeit realer Entwicklungen auf Finanzmärkten.[43]

Der Begriff "Manager-Sentiment“ bezeichnet die emotionalen und psychologischen Haltungen, die Manager in ihren Entscheidungen und Äußerungen zum Ausdruck bringen.[44] Diese können auch einen Verhaltensbias umfassen, der sich in irrationalen Überzeugungen über die zukünftige Aussicht ihrer Unternehmen manifestiert.[45] Im Rahmen ihrer Tätigkeit sammeln Manager private Informationen über die zukünftig erwarteten Cashflows und Risiken ihres Unternehmens.[46] Diese bleiben in den quantitativen Unternehmensberichten allerdings weitgehend unberücksichtigt, da solche Berichte primär vergangenheitsbezogene und standardisierte Daten enthalten, die nur geringe Einblicke in die internen Einschätzungen und strategischen Überlegungen des Managements bieten. Dahingegen können qualitative Offenlegungen den Anlegern zumindest teilweise dabei behilflich sein, das Manager-Sentiment und die damit verbundenen Insiderinformationen der Unternehmensleitung zu identifizieren und diese Informationen in ihre Anlageentscheidungen einfließen zu lassen. Earnings Calls stellen hierbei eine zentrale Form der Offenlegung dar, in denen subjektive Wahrnehmungen und emotionale Reaktionen zum Ausdruck kommen, die das Sentiment der Manager widerspiegeln.[47] Die präsentierten verbalen und nonverbalen Hinweise der Unternehmensführung können als Indikatoren für zukünftige Geschäftserwartungen betrachtet werden, die sich nicht unmittelbar aus den reinen Zahlen ableiten lassen. Daher wird folgende Hypothese aufgestellt:

> **H1a:** *Ein positives Sentiment in Earnings Calls führt zu signifikant positiven Marktreaktionen, während ein negatives Sentiment signifikant negative Marktreaktionen hervorruft*

Sowohl das Sentiment von Managern als auch das von Investoren stellt einen wesentlichen Faktor in der Dynamik der Finanzmärkte dar.[48] Dennoch lassen sich Unterschiede in den Hauptinteressen und Einflussfaktoren feststellen. Investoren fokussieren sich primär auf die erwarteten Renditen, Wachstumsaussichten und Risiken in ihren Portfolios. Ihr Sentiment umfasst die kollektiven emotionalen und psychologischen Haltungen gegenüber einem Unternehmen oder dem Markt im Allgemeinen, die nicht durch wirtschaftliche Fundamentaldaten gestützt sein müssen. Im Gegensatz dazu ist die Unternehmensführung primär an den zukünftigen Aussichten der von ihnen betreuten Unternehmen interessiert. Gregory (2021) konnte nachweisen, dass die Prognosefähigkeit des Manager-Sentiments in erster Linie auf privaten Informationen basiert, während die Investorenstimmung überwiegend durch öffentliche Informationen geprägt ist.[49] Dies zeigt, dass beide Gruppen über unterschiedliche Informationsgrundlagen verfügen. Darüber hinaus stellten Jiang et al. (2019) fest, dass das Sentiment von Managern und Investoren nicht in gegenseitiger Abhängigkeit zueinandersteht, sondern unterschiedliche Aspekte des Marktsentiments erfasst und sich ergänzt.[50] Sie betonen, dass das Manager-Sentiment stark durch interne Erwartungen getrieben wird, die häufig übermäßig optimistisch sind.

Empirische Evidenz sowie theoretische Überlegungen legen nahe, dass insbesondere Aktien, deren Bewertung schwierig ist, deren Handel kostenintensiv ist oder die einen spekulativen Charakter besitzen, eine höhere Anfälligkeit für Sentiment aufweisen.[51] Die Interpretation von Informationen gestaltet sich für Investoren mit unzureichendem Informationsstand als Herausforderung. Dies kann dazu führen, dass vermehrt auf alternative Informationsquellen zurückgegriffen wird, um die bestehende Informationsasymmetrie zu reduzieren. Auch bei der Untersuchung von Earnings Calls wird ersichtlich, dass Informationsasymmetrien zu einer verstärkten Beeinflussung durch Sentiment führen. Bochkay et al. (2020) zeigten beispielsweise, dass bei Unternehmen mit einem weniger transparenten Informationsumfeld sowohl das Handelsvolumen als auch die Kursreaktionen während einer Telefonkonferenz ansteigen, wenn sich die Manager einer extremeren Sprache bedienen.[52] Auf Basis dieser Erkenntnisse lässt sich folgende Hypothese formulieren:

> **H1b:** *Je geringer die Informationslage der Investoren eines Unternehmens, desto stärker reagieren die Investoren auf das Sentiment in Earnings Calls*

Diese Differenzierung in der Wahrnehmung des Sentiments führt zu der Frage, wie Manager durch gezielte Offenlegungen und Sprachstrategien versuchen, Informationsasymmetrien zu verringern und die Erwartungen der Investoren zu steuern. Gemäß der Signaling-Theorie nutzen Manager freiwillige Offenlegungen und den Sprachstil ihrer Aussagen, um die Erwartungen der Investoren bezüglich künftiger Unternehmensleistungen mit ihren eigenen Prognosen in Übereinstimmung zu bringen.[53] Diese Strategie zielt darauf

[42] Vgl. Todd et al. (2024, p. 2).
[43] Vgl. M. Baker und Wurgler (2007, p. 130).
[44] Vgl. Zhou (2018, pp. 241f).
[45] Vgl. Collins et al. (2024, pp. 2 und 8).
[46] Vgl. Arslan-Ayaydin et al. (2016, pp. 133f).
[47] Vgl. De Amicis et al. (2021); Blau et al. (2015, p. 206).
[48] Vgl. Collins et al. (2024, pp. 7f).
[49] Vgl. Gregory (2021).
[50] Vgl. Jiang et al. (2019, p. 148).
[51] Vgl. Jiang et al. (2019, p. 128).
[52] Vgl. Bochkay et al. (2019, p. 56).
[53] Vgl. Pouryousof et al. (2023); Ajinkya und Gift (1984, p. 443).

ab, Unterschiede im Informationsstand zu reduzieren und die Glaubwürdigkeit zu stärken, um dadurch die Marktreaktionen optimal zu steuern.[54] Die Formulierung der Erzählungen gewährt den Investoren Einblick in die Erwartungen der Manager bezüglich künftiger Entwicklungen des Unternehmens. In der wissenschaftlichen Literatur konnte ein positiver Zusammenhang zwischen einer positiven Ausdrucksweise und Leistungskennzahlen, wie beispielsweise Gewinnen und Cashflows, festgestellt werden. Eine übermäßig positive Darstellung der Manager in narrativen Offenlegungen wird durch verschiedene Faktoren limitiert.[55] Ein wesentlicher Kostenfaktor ist dabei der potenzielle Verlust an Glaubwürdigkeit, die für Investoren von entscheidender Bedeutung ist. Zudem kann Optimismus in finanziellen Offenlegungen als Hinweis auf eine hohe Branchenrentabilität interpretiert werden, was potenzielle Wettbewerber dazu ermutigen könnte, in die Branche einzutreten. Schließlich birgt die Möglichkeit potenzieller Rechtsstreitigkeiten, die durch übermäßig positive Angaben in qualitativen Offenlegungen hervorgerufen werden können, ein bedeutendes Risiko.[56]

Die Betrachtung der Ausdrucksweise von Unternehmensvertretern in Earnings Calls als objektiver Indikator für private Informationen von Managern über zukünftige Unternehmensleistungen vernachlässigt jedoch die Tatsache, dass Manager von einer Vielzahl von Anreizen und Vorurteilen beeinflusst werden können, die die Transparenz der Offenlegung beeinträchtigen können.[57] In der Konsequenz kann es zu einer irreführenden Wahrnehmung seitens der Anleger kommen.[58] Der Agency-Theorie zufolge treffen Manager mit Insiderwissen häufig Entscheidungen, die ihre eigenen Interessen mehr berücksichtigen als die der Aktionäre.[59] Opportunistische Manager tendieren demnach dazu, ihre Schwächen zu verbergen und ihre Erfolge übermäßig hervorzuheben, um positive Berichterstattung zu maximieren und negative Konsequenzen zu minimieren.[60] Eine negative Stimmung unter Investoren sowie ungünstige Marktreaktionen können Manager in eine Drucksituation bringen, in der persönliche Faktoren wie Vergütung, Karriereentwicklung und Erhalt des eigenen Rufs eine ausschlaggebende Rolle spielen.[61] Beispielsweise legten Arslan-Ayaydin et al. (2016) dar, dass Manager mit aktienbasierter Vergütung eine positivere Tonalität in Gewinnmitteilungen wählen, um über steigende Kurse ihre Vergütung zu erhöhen.[62] Huang et al. (2014) stellten fest, dass Unternehmen, die im Kontext spezifischer Ereignisse wie beispielsweise der Ausgabe neuer Aktien oder Fusionen eine positive Wahrnehmung bei Investoren anstreben, dazu tendieren, einen optimistischen Tonfall in der entsprechenden Unternehmenskommunikation zu verwenden.[63] Es besteht also die Möglichkeit, dass Anleger die in nicht-finanziellen Berichten ausgedrückte Stimmung der Unternehmensführung aufgreifen, auch wenn diese nicht den tatsächlichen Wertgrundlagen des Unternehmens entspricht.[64] Eine übermäßig positive (negative) Berichterstattung des Managements könnte folglich zu einer spekulativen Überbewertung (Unterbewertung) des Marktes führen.

Darüber hinaus wird von der Verhaltenstheorie angenommen, dass eine gewisse Inkonsistenz im Tonfall von Managern entweder zufällige oder persönliche Ursachen hat.[65] Davis et al. (2015) demonstrierten, dass Manager-spezifisch optimistische (oder pessimistische) Tendenzen den Sprachstil der finanziellen Offenlegungen über die aktuelle oder zukünftige finanzielle Unternehmensleistung und strategische Anreize der Manager hinaus beeinflussen.[66] Des Weiteren lässt sich feststellen, dass bestimmte beobachtbare Faktoren, wie frühe berufliche Erfahrungen und ehrenamtliches Engagement, einen Einfluss auf die Art und Weise haben, wie Manager kommunizieren. So konnten De Amicis et al. (2021) und Klevak et al. (2024) geschlechtsspezifische Unterschiede im Tonfall nachweisen.[67]

In einigen Forschungsarbeiten wird folglich die These aufgestellt, dass Nuancen im Tonfall der Bekanntgabe von Unternehmensmeldungen lediglich begrenzte, wenn überhaupt messbare, Auswirkungen auf die langfristige Kursentwicklung haben.[68] In diesem Zusammenhang wird das in Unternehmensveröffentlichungen zum Ausdruck gebrachte Sentiment von mehreren Autoren als „leeres Gerede“ bezeichnet. Daraus lässt sich die folgende Hypothese ableiten:

> **H1c:** *Die initale Marktreaktion auf das Sentiment in Earnings Calls erfährt über einen längeren Zeitraum hinweg eine Korrektur*

Vor diesem Hintergrund ist es ebenfalls denkbar, dass die Präsentation sowie die anschließende Fragerunde unterschiedliche Dimensionen des Management-Sentiments widerspiegeln, die jeweils separat die Marktreaktionen beeinflussen können.[69] Das spezifische Sentiment in dem spontaneren Q&A-Abschnitt bietet möglicherweise detailliertere und authentischere Einblicke in die wahre Unternehmenslage als die strategisch vorformulierte Darstellung in der Präsentation. Da Manager im Q&A auf kritische und unerwartete Fragen reagieren müssen, könnten etwaige Schwächen oder Unsicherheiten deutlicher werden und die persönliche Haltung sowie der emotionale Zustand des Managements unmittelbarer reflektiert werden. Daraus ergeben sich die folgenden zwei Hypothesen:

[54] Vgl. Rahman (2023).
[55] Vgl. Arslan-Ayaydin et al. (2016, p. 134).
[56] Vgl. Rogers et al. (2011, p. 2179).
[57] Vgl. Arslan-Ayaydin et al. (2016, pp. 133f).; Beyer et al. (2010, pp. 196f).
[58] Vgl. Pouryousof et al. (2023).
[59] Vgl. Jensen und Meckling (1976, pp. 308f. and 312f.)
[60] Vgl. Rahman (2023); Mather et al. (2021, p. 3); Matsumoto et al. (2011, p. 230).
[61] Vgl. R. Luo und Ye (2023, p. 837).
[62] Vgl. Arslan-Ayaydin et al. (2016, p. 133).
[63] Vgl. Huang et al. (2013, p. 1111).
[64] Vgl. Gupta (2019, p. 17); Jiang et al. (2019, p. 127).
[65] Vgl. Pouryousof et al. (2023).
[66] Vgl. Davis et al. (2014, p. 671).
[67] Vgl. Klevak et al. (2024); De Amicis et al. (2021).
[68] Vgl. Hennig et al. (2023).
[69] Vgl. Brockman et al. (2015, pp. 24f).

H2a: *Das Sentiment der Präsentation und das Sentiment im Q&A-Abschnitt hat jeweils einen signifikanten und unabhängigen Einfluss auf die Marktreaktion*

H2b: *Das Sentiment des Q&A-Abschnitts hat einen stärkeren Einfluss auf die Markreaktion als das Sentiment der Präsentation.*

Neben der Unternehmensführung nehmen auch Analysten eine bedeutsame Rolle in Earnings Calls ein. Sie verfügen über einen umfassenderen Zugang zu Informationsquellen als durchschnittliche Investoren und spielen eine entscheidende Rolle als Vermittler von Unternehmensinformationen.[70] Im Vergleich zu Managern verfügen Analysten über einen Anreizvorteil, da ihre Interessen in höherem Maße mit jenen der externen Investoren übereinstimmen.[71] Während Manager dazu neigen, Informationen in einem optimistischen Licht darzustellen, nehmen Analysten eine kritischere Haltung ein und legen ein stärkeres Augenmerk auf potenzielle Risiken. Die Ausdrucksweise von Analysten unterliegt jedoch auch bestimmten Einschränkungen. Unternehmen neigen dazu, überwiegend wohlgesonnene Analysten einzuladen.[72] Zudem lässt sich feststellen, dass Analysten ihren Kommunikationsstil strategisch einsetzen, um eine positive Beziehung zur Unternehmensführung zu etablieren und ihre Informationsgrundlage zu sichern.[73] Zur Untersuchung der Unterschiede zwischen Manager- und Analysten-Sentiment werden die Hypothesen formuliert:

H2c: *Das Sentiment der Manager und Analysten innerhalb des Q&A-Abschnitts hat jeweils einen signifikanten und unabhängigen Einfluss auf die Marktreaktion*

H2d: *Das Sentiment der Fragen von Analysten hat einen stärkeren Einfluss auf die Marktreaktion als das Sentiment der Antworten von Managern innerhalb des Q&As*

2.3 Sentimentextraktion mittels textueller Analysen

Die gewonnenen Erkenntnisse veranschaulichen die Komplexität der Informationsübermittlung in Unternehmensoffenlegungen und verdeutlichen die Herausforderungen für Investoren, authentische Signale korrekt zu interpretieren. Vor diesem Hintergrund erweist sich die Sentimentanalyse als ein entscheidendes Instrument, um die psychologischen Einflussfaktoren auf die Märkte zu quantifizieren und zu analysieren.[74] Es ist daher von entscheidender Bedeutung, zu verstehen, wie Sentiment am besten erfasst werden kann. Zur Ermittlung von Sentiment im Kontext finanzieller Analysen werden im Wesentlichen drei Ansätze verfolgt: die Beobachtung von Marktdaten, die Durchführung von Umfragen und die Auswertung von Texten. Kritikpunkte an marktbasierten Ansätzen sind die begrenzte Datenverfügbarkeit, der vergangenheitsbezogene Charakter und Endogenitätsprobleme.[75] Umfragen bieten zwar wertvolle Einblicke in den Meinungsbildungsprozess der Marktteilnehmer, sind jedoch mit einem hohen Aufwand bei der Durchführung verbunden. Des Weiteren ist die Datenverfügbarkeit hinsichtlich des Umfangs und der Häufigkeit eingeschränkt und die Ergebnisse hängen im hohem Maße vom Erhebungsdesign ab, wobei die Gefahr von Antwortverzerrungen besteht.[76] Mit Entwicklungen in der Rechenleistung, der Computerlinguistik und der Verfügbarkeit von Daten hat daher die Extraktion von Sentiment aus Texten an Bedeutung gewonnen. Dies resultierte in einer verstärkten Untersuchung linguistischer Signale in der Fachliteratur für Rechnungswesen und Finanzen. Die textuelle Analyse erlaubt die Extraktion sprachlicher Informationen, die für herkömmliche Leser oft nicht offensichtlich sind.[77] Dabei geht die textbasierte Sentimentanalyse über die bloße Inhaltsbewertung hinaus, indem sie darauf abzielt, die emotionale Haltung oder persönliche Meinung hinter den Texten zu erfassen und zu quantifizieren. Textuelle Auswertungen schneiden bei der Aktienmarktprognose besser ab als markt- und umfragebasierte Methoden.[78] Dies deutet darauf hin, dass sie in diesem Bereich bislang unzureichend beachtet oder berücksichtigt werden.

Die Wechselwirkung zwischen textuellem und marktteilnehmerbezogenem Sentiment ist dabei komplex und noch nicht vollständig erforscht, einschließlich der Art und Weise, wie Adressaten das textuelle Sentiment deuten.[79] Zur Untersuchung des Sentiments erfolgt vielfach eine Analyse des „Tons“ von Texten, wobei dieser sowohl vom Inhalt als auch von der Wortwahl abhängt.[80] Ein positiver Ton kann demnach durch die Hervorhebung positiver Ergebnisse und/oder einer positiven Berichterstattung über die Ergebnisse erreicht werden. Ein gängiges Verfahren zur Bestimmung des Tons ist die Ermittlung der Anzahl positiver und negativer Wörter. Dabei lässt ein überwiegender Anteil positiver Wörter im Vergleich zu negativen auf einen optimistischeren Ton schließen. Der Begriff des textuellen Sentiments ist jedoch nicht allein auf die Dimensionen Positivität und Negativität beschränkt, sondern kann auch auf andere, darüberhinausgehende Dimensionen anwendbar sein, wie etwa stark-schwach und aktiv-passiv.[81] Im weiteren Verlauf dieser Untersuchung wird der Ton eines Textes, gemessen an der relativen Häufigkeit positiver und negativer Wörter, als Indikator für das vorherrschende Sentiment herangezogen.

[70] Vgl. Y. Chen et al. (2023).
[71] Vgl. Brockman et al. (2015, p. 25).
[72] Vgl. R. Luo und Ye (2023, p. 834); Lee (2015, p. 229)
[73] Vgl. Milian et al. (2017, pp. 14f.); Brockman et al. (2015, p. 25).
[74] Vgl. Zhou (2018, pp. 204–242).
[75] Vgl. Bai et al. (2023).
[76] Vgl. Zhou (2018, pp. 248–250).
[77] Vgl. Pennebaker et al. (2003, p. 554).
[78] Vgl. Zhou (2018, p. 250).
[79] Vgl. Kearney und Liu (2014, p. 172).
[80] Vgl. Bassyouny et al. (2022); Henry (2008, p. 377).
[81] Vgl. Kearney und Liu (2014, p. 172).

Die Anfänge der Textanalyse in der Finanzforschung demonstrieren das Potenzial der Untersuchung von Marktbewegungen unter Einbezug von Sentiment aus verschiedenen Textquellen. In einer der ersten Anwendungen der Textanalyse im Jahr 2004 untersuchten Antweiler und Frank den Einfluss von Börsenforen auf den Aktienmarkt.[82] Sie stellten fest, dass das Nachrichtenvolumen in Börsenforen in der Lage ist, die Volatilität des Marktes vorherzusagen. In Anlehnung an den Ton der Nachrichten in diesen Foren konnten Das und Chen (2007) nachweisen, dass die Entwicklung von Indexrenditen prognostizierbar ist.[83] Im selben Jahr zeigte Tetlock, dass das in einer Kolumne des Wall Street Journal ausgedrückte Sentiment die zukünftige Entwicklung der Aktienkurse vorhersagen kann.[84] In der Folge wurde in der Literatur ein breites Spektrum unterschiedlicher Textsorten analysiert, wie Nachrichtenartikel,[85] soziale Medien[86] und Unternehmensmitteilungen.[87]

Insbesondere die Analyse des in Unternehmensveröffentlichungen geäußerten Sentiments erweist sich als hilfreich, um die Bedeutung qualitativer Informationen für die Bewertung einzelner Unternehmen und deren Kapitalmarktperformance besser zu verstehen.[88] Obwohl die finanzielle Berichterstattung oft mit Zahlen in Verbindung gebracht wird, ist letztlich Sprache das primäre Kommunikationsmedium für die Vermittlung vergangener und zukünftiger Unternehmensleistungen.[89] Die unzureichende Regulierung der Tonalität in der Finanzberichterstattung eröffnet der Unternehmensleitung einen größeren Spielraum bei der Hervorhebung bestimmter Aspekte der Unternehmensleistung sowie der Wortwahl zur Beschreibung dieser Leistung.[90] Es besteht die Möglichkeit, dass zwei Unternehmen, deren Finanzkennzahlen identisch sind, Offenlegungen herausgeben, die eine sehr unterschiedliche Gesamtbotschaft bezüglich der Unternehmensleistung vermitteln. Der Großteil qualitativer Informationen über Unternehmen ist öffentlich zugänglich und in Textform verfügbar.[91] Im Rahmen der Untersuchung von Unternehmensmitteilungen wird insbesondere analysiert, wie optimistische und pessimistische Sprache von Managern in narrativen Berichten eingesetzt wird, um wesentliche Informationen über ein Unternehmen zu kommunizieren.[92] Dabei wurden Textanalysen verschiedener Arten von Unternehmensveröffentlichungen durchgeführt, darunter Pressemitteilungen,[93] Gewinnbekanntgaben[94] und Managementkommentare in den Jahres- und Quartalsberichten.[95]

Insbesondere im Kontext von Earnings Calls eröffnet die textuelle Analyse die Möglichkeit, die Einschätzungen und Bewertungen der Teilnehmer in einer systematischen Weise zu erfassen und zu evaluieren. Computer können in diesen schriftlichen Aufzeichnungen Nuancen im Tonfall der Unternehmensleitung identifizieren, die von Analysten möglicherweise übersehen werden.[96] Die Anwendung von Computeralgorithmen zur Sentimentanalyse kann demzufolge zu einer präziseren und umfassenderen Erfassung der kommunizierten Informationen führen. Da es keine gesetzliche Verpflichtung gibt, schriftliche Transkripte von Earnings Calls bereitzustellen, sind nicht alle Telefonkonferenzen dokumentiert.[97] Dennoch werden sie von einer steigenden Anzahl an Unternehmen oder Drittanbietern transkribiert. Das Fehlen solcher Dokumente kann insbesondere für Einzel- und nichtprofessionelle Investoren von Nachteil sein.[98] Obwohl die meisten Unternehmen ihre Earnings Calls zeitnah nach Veröffentlichung der Ergebnisse über Webcasts oder Einwahldienste live zugänglich machen, sind die enthaltenden Informationen oft auch über den Termin hinaus relevant. Da mehrere Konferenzen gleichzeitig stattfinden können, ist es für Anleger oft schwierig, alle in Echtzeit zu verfolgen. Schriftliche Transkripte erlauben zudem eine erleichterte Verarbeitung der Informationen, da sie eine wiederholbare und flexible Analyse ermöglichen, während die Verarbeitung von Informationen in Live-Calls in Echtzeit erfolgen muss und stimmliche sowie visuelle Ablenkungen zusätzliche kognitive Belastungen mit sich bringen.[99] Vor diesem Hintergrund erweisen sich Transkripte als effiziente Alternative, die für zahlreiche Anleger einen wesentlichen Bestandteil ihrer Anlageforschung darstellen.

Die Literatur zum Einfluss von Sentiment in Earnings-Call-Transkripten auf die Kapitalmärkte bietet vielfältige Einblicke in die Dynamik zwischen Unternehmenskommunikation und Marktreaktionen. In einer Studie von Price et al. (2012) wurde erstmals der Zusammenhang zwischen dem Sentiment in diesen Telefonkonferenzen und Kapitalmarktreaktionen untersucht.[100] Die Ergebnisse zeigen, dass das Sentiment in Earnings Calls einen signifikanten Einfluss auf verschiedene Finanzmarktgrößen ausübt. Dazu zählen abnormale Renditen sowohl während des Gewinnankündigungsfensters als auch über einen längeren Zeitraum hinweg sowie das Handelsvolumen. Borochin et al. (2017) ergänzten diese Erkenntnisse, indem sie feststellten, dass pessimistische Stimmungen während Earnings Calls zu größerer Marktunsicherheit führen als optimistische Stimmungen.[101] Insbesondere ist der starke Einfluss einer negativen Analystenstimmung während des Q&A-Abschnitts hervorzuheben. Dies

[82] Vgl. Antweiler und Frank (2004, p. 1292).
[83] Vgl. Das und Chen (2007, p. 1386).
[84] Vgl. Tetlock (2007, p. 1166).
[85] Vgl. Bai et al. (2023, p. 8); Tetlock (2007, p. 1166).
[86] Vgl. Agrawal et al. (2018, p. 86); Renault (2017, p. 37).
[87] Vgl. Jiang et al. (2019, p. 148); Loughran und McDonald (2011, p. 61).
[88] Vgl. Kearney und Liu (2014, p. 173).
[89] Vgl. Hales et al. (2011, p. 224).
[90] Vgl. Henry und Leone (2016, p. 153).
[91] Vgl. Kearney und Liu (2014, p. 181).
[92] Vgl. Bassyouny et al. (2022).
[93] Vgl. Arslan-Ayaydin et al. (2016, p. 133); Huang et al. (2013, p. 1111).
[94] Vgl. DeBoskey et al. (2018, p. 658); Y. Luo und Zhou (2017, p. 454); Tama-Sweet (2014, p. 276).
[95] Vgl. Yan et al. (2023, p. 2); D'Augusta und DeAngelis (2019, p. 133).
[96] Vgl. Loughran und McDonald (2016, p. 1188).
[97] Vgl. Elliott et al. (2023, pp. 9 und 12).
[98] Vgl. Altman (2019).
[99] Vgl. Elliott et al. (2023, pp. 9 und 12).
[100] Vgl. Price et al. (2012, p. 1006).
[101] Vgl. Borochin et al. (2018, p. 76).

führen die Autoren darauf zurück, dass Analysten als vertrauenswürdigere Informationsquellen gelten als Unternehmensvertreter. Fu et al. (2019) stellten fest, dass ein höherer Optimismus in Earnings Calls das Risiko eines Aktienkursrückgangs signifikant reduzieren kann.[102] Auch sie identifizierten den Q&A-Teil der Calls als besonders informativ, was darauf hindeutet, dass die Interaktion zwischen Analysten und Managern entscheidend für die besonderen Marktreaktionen auf Earnings Calls ist.

In der wissenschaftlichen Literatur wird vielfach die Auffassung vertreten, dass das Sentiment der Analysten im Q&A-Abschnitt von Earnings Calls eine größere Relevanz besitzt als das der Manager. Brockman et al. (2015) zeigten, dass Manager in Earnings Calls eine tendenziell optimistischere Haltung als andere Teilnehmer einnehmen.[103] Investoren vertrauen in erster Linie dem Ton von Analysten, die ähnliche Anreize wie sie haben, obwohl Manager über einen erheblichen Informationsvorteil verfügen. Ergänzend stellten X. Chen und Tian (2018) fest, dass Analysten Manager dazu bringen können, ihren anfänglichen Optimismus aufzugeben, und dass Aktienkurse stärker auf den Ton der Analysten als auf den der Manager reagieren.[104] Im Gegensatz dazu wiesen Milian et al. (2017) darauf hin, dass der Ton des gesamten Earnings Calls Aktienrenditen vorhersagen kann, nicht jedoch der der Analysten.[105] Auch Yang et al. (2023) zeigten, dass die Aussagen der Manager im Q&A-Teil des Earnings Calls für den Markt bedeutsamer sind als die der Analysten.[106] Diese widersprüchlichen Erkenntnisse verdeutlichen, dass der Einfluss von Sentiment in verschiedenen Segmenten von Earnings Calls auf Investoren noch nicht vollständig erforscht ist. Die unterschiedlichen Ergebnisse könnten insbesondere auf Variationen in der Informationslage der Anleger sowie auf methodische Unterschiede in den Studien zurückzuführen sein. Weitere Forschung ist erforderlich, um diese Diskrepanzen zu klären und ein umfassenderes Verständnis dafür zu entwickeln, wie und warum Sentiment in verschiedenen Segmenten von Earnings Calls unterschiedliche Auswirkungen auf den Markt hat.

2.4 Lexikonbasierte Textauswertung

Die in der Literaturübersicht dargestellten Beobachtungen verdeutlichen, dass aufgrund der inhärenten Ungenauigkeiten dieses Vorhabens die Methodik von zentraler Bedeutung ist, mit der qualitative Stimmungen, Einstellungen und Emotionen in Texten quantifiziert und in ein Sentimentmaß umgewandelt werden.[107] Während quantitative Forschung auf etablierten Normen und klar definierten Maßeinheiten basiert, erfordert die Auswertung von Texten aufgrund der Mehrdeutigkeit sprachlicher Zeichen eine kontextabhängige Interpretation. In der frühen Kapitalmarktforschung wurden subjektive Bewertungen des textuellen Tons durch menschliche Codierer vorgenommen.[108] Heute werden nahezu ausschließlich computergestützte Programme zur Verarbeitung natürlicher Sprache eingesetzt, um Indikatoren für das Sentiment der Offenlegung zu ermitteln. Diese Verfahren lassen sich der lexikonbasierten Analyse oder dem maschinellen Lernen zuordnen.

Der prominenteste Ansatz zur Ermittlung der Stimmung eines Textes besteht darin, den Text in seine Wörter zu zerlegen und diese anschließend mithilfe eines Lexikons zu kategorisieren.[109] Für diese lexikonbasierten Textanalysen wurden vordefinierte Wörterbücher entwickelt, die Wörtern Attribute wie „positiv" oder „negativ" zuweisen. Diese Methodik lässt sich dem sogenannten „Bag of Words"-Ansatz zuordnen. Dabei wird die Häufigkeit von Wörtern innerhalb eines Textdokumentes untersucht, ohne die Reihenfolge und den unmittelbaren Zusammenhang der Wörter zu berücksichtigen. Der lexikonbasierte Ansatz kann demnach Schwierigkeiten aufweisen, den Kontext und die Mehrdeutigkeit von Wörtern adäquat zu erfassen.[110] Zudem besteht die Möglichkeit, dass Lexika veraltet sind und Änderungen in der Sprache nicht angemessen widerspiegeln.[111]

Technologische Fortschritte haben in den letzten Jahren die Möglichkeiten, maschinelles Lernen in der Finanzanalyse einzusetzen, erheblich erweitert und bieten somit eine wertvolle Ergänzung zu den traditionellen, lexikonbasierten Ansätzen.[112] Der Begriff des maschinellen Lernens bezeichnet die Fähigkeit von Computern, aus Daten Muster zu erkennen und Vorhersagen zu treffen.[113] Die Anwendung von Algorithmen und statistischen Modellen erlaubt die Übertragung der gewonnenen Erkenntnisse auf andere Daten. Zu den Vorteilen maschinellen Lernens im Kontext textueller Analysen gehören demnach die Fähigkeit, sich an Veränderungen anzupassen sowie eine potenzielle höhere Genauigkeit bei spezifischen Aufgaben. Ein Beispiel für den Einsatz von maschinellem Lernen liefert die Studie von Frankel et al. (2021).[114] In dieser wurden verschiedene Algorithmen, darunter neuronale Netze, eingesetzt, um Klassifikationsmodelle zu trainieren, die in der Lage sind, das Sentiment in Jahresberichten und Transkripten von Earnings Calls zu erkennen. Die Resultate demonstrieren, dass maschinelle Lernmodelle eine höhere Präzision aufweisen als konventionelle, auf Lexika basierende Ansätze. Die von Renault (2020) durchgeführte Analyse verdeutlicht jedoch, dass komplexe Algorithmen nicht zwangsläufig präzisere Ergebnisse liefern als lexikonbasierte Methoden.[115]

Vor diesem Hintergrund erweist sich der lexikalische Ansatz, der im Rahmen der empirischen Untersuchung An-

[102] Vgl. Fu et al. (2019, p. 644).
[103] Vgl. Brockman et al. (2015, p. 40).
[104] Vgl. X. Chen und Tian (2018, p. 1348).
[105] Vgl. Milian et al. (2017, p. 73).
[106] Vgl. Yang et al. (2022, p. 2177).
[107] Vgl. Loughran und McDonald (2016, pp. 1189–1191).
[108] Vgl. Henry und Leone (2016, p. 156).
[109] Vgl. Loughran und McDonald (2016, pp. 1198–1200).
[110] Vgl. Loughran und McDonald (2016, p. 1226).
[111] Vgl. Frankel et al. (2021, p. 5516).
[112] Vgl. Todd et al. (2024, p. 4).
[113] Vgl. Frankel et al. (2021, pp. 5517f).
[114] Vgl. Frankel et al. (2021, pp. 5514f).
[115] Vgl. Renault (2017, p. 31).

wendung findet, weiterhin als eine bewährte Methode, die gegenüber maschinellen Lernansätzen eine Reihe von Vorteilen aufweist. Die Verwendung einer vordefinierten Wortliste im Rahmen eines lexikonbasierten Ansatzes gewährleistet eine einfache Implementierung sowie Reproduzierbarkeit der Forschungsergebnisse.[116] Im Gegensatz zu maschinellen Lernverfahren benötigen lexikalische Analysen keine umfangreichen Trainingsdaten und ermöglichen eine schnelle Implementierung. Nach Festlegung des zugrunde liegenden Wörterbuchs sind die Ergebnisse objektiv und robust gegen Überanpassung. Maschinelle Lernverfahren werden häufig als Blackbox wahrgenommen, während der lexikonbasierte Ansatz eine transparente und nachvollziehbare Analyse der Ergebnisse ermöglicht. Das Verfahren eignet sich insbesondere für die Analyse großer Datenmengen, da die computergestützte Häufigkeitszählung der Wörter vergleichsweise geringe Rechenressourcen erfordert. Lexika, die gezielt auf die Eigenschaften des vorliegenden Textes abgestimmt sind, liefern bei der Sentiment-Extraktion Ergebnisse, die häufig mit denen differenzierterer Instrumente wie maschinellen Lernens vergleichbar sind.[117] Die Auswahl der Wortliste, die als Grundlage für die Klassifizierung von Wörtern dient, erweist sich bei einem lexikonbasierten Ansatz als kritischer Faktor.[118] Zur Prüfung der Effekte des ausgewählten Lexikons auf das ermittelte Sentiment wird folgende Hypothese aufgestellt:

> **H3a:** *Der Zusammenhang zwischen Sentiment in Earnings Calls und der Marktreaktion variiert je nach dem verwendeten Sentiment-Lexikon*

Zu Beginn der textbasierten Analysen in den Bereichen Rechnungswesen und Finanzen wurde auf allgemeine Wortlisten wie das Harvard-GI-Lexikon zurückgegriffen, da dieses früh verfügbar war und zu diesem Zeitpunkt noch keine spezielleren Listen für diese Bereiche existierten.[119] Stone und Hunt entwickelten im Jahr 1961 das Lexikon für das General Inquirer System.[120] Dieses computergestützte Analysewerkzeug verwendet vordefinierte Wortlisten, um psychologische und soziale Phänomene in Texten zu erfassen und zu quantifizieren. Das Harvard-GI-Lexikon enthält 2.005 negative und 1.637 positive Wörter, wobei auch verschiedene Wortformen berücksichtigt werden.[121]

Die Anwendung generischer Lexika zur Erfassung finanziellen Sentiments kann zu Ungenauigkeiten führen, da sie nicht speziell für den Kontext der Finanzwirtschaft entwickelt wurden.[122] Um die spezifischen Nuancen des Finanzwesens angemessen erfassen zu können, ist es erforderlich, die Wörterbücher kontextspezifisch anzupassen oder zu erweitern. Loughran und McDonald (2011) berichteten, dass nahezu 75 % der als Harvard-GI-negativ klassifizierten Wörter in Finanzdokumenten keine negative Bedeutung aufweisen.[123] So wird beispielsweise das Wort „beat" (schlagen) im GI-Wörterbuch als negativ eingestuft, obwohl es in Gewinnmitteilungen häufig verwendet wird, um das Übertreffen von Erwartungen zu beschreiben.[124] Auch im Hinblick auf die Analyse von Earnings Calls hat sich die Verwendung kontextspezifischer Wörterbücher als überlegen erwiesen.[125] In ihrer Studie aus dem Jahr 2012 konnten Price et al. nachweisen, dass das finanzspezifische Henry-Wörterbuch im Vergleich zum generischen Harvard-GI eine präzisere Methode zur Ermittlung eines Tons aus Earnings Calls darstellt, der Rückschlüsse auf Aktienrenditen zulässt.[126] Auch Davis et al. (2015) wiesen darauf hin, dass die kontextspezifischen Henry- und LM-Wörterbücher ein generisches Lexikon bei der Bestimmung des Tons in Earnings Calls übertreffen.[127]

Die erste speziell für finanzielle Texte entwickelte Wortliste wurde im Jahr 2008 von Henry erstellt.[128] Sie basiert auf der Analyse von Gewinnmitteilungen der Telekommunikations- und Computerdienstleistungsbranche der Jahre 1998 bis 2002. Im Vergleich zu generischen Wörterbüchern wie GI umfasst das Henry-Lexikon nur eine begrenzte Anzahl an Wörtern, nämlich 105 positive und 85 negative.[129] In der Konsequenz findet eine Vielzahl von im Bereich der Geschäftskommunikation häufig verwendeten Wörtern keine Berücksichtigung, was eine Beeinträchtigung der Erfassung der für Finanztexte relevanten Attribute und Töne zur Folge haben könnte. Um Sentiment in Geschäftsberichten und Finanzdokumenten besser bewerten zu können, erstellten Loughran und McDonald daher das LM-Wörterbuch.[130] Das Lexikon umfasst die Kategorien negativ, positiv, unsicher, streitlustig, stark modal und schwach modal. Die Zuordnung der Wörter erfolgte durch die Analyse von US-Jahresberichten aus den Jahren 1994 bis 2008, mit dem Ziel, eine umfassende Liste von Wörtern zu erstellen, die nicht gezielt von Unternehmen vermieden werden können. Das LM-Lexikon wird in einem zweijährigen Rhythmus aktualisiert und umfasst zum Zeitpunkt der Erstellung dieser Arbeit 2.355 negative und 354 positive Wörter. Im Vergleich zu anderen, in der Rechnungswesen- und Finanzliteratur verbreiteten Wortwörterbüchern weist es demnach zwei wesentliche Vorteile auf: Einerseits wurde es speziell für die Finanzkommunikation erstellt und basiert auf der wahrscheinlichsten Interpretation eines Wortes im Geschäftskontext, andererseits zeichnet es sich durch eine breite Abdeckung von Wörtern aus. Daher wird die folgende Hypothese getestet:

[116] Vgl. Frankel et al. (2021, pp. 5516f).; Loughran und McDonald (2016, p. 1200).
[117] Vgl. Renault (2017, p. 37); Henry und Leone (2016, p. 174).
[118] Vgl. Loughran und McDonald (2011, p. 36).
[119] Vgl. Loughran und McDonald (2016, p. 1201).
[120] Vgl. Stone und Hunt (1963, pp. 241f).
[121] Vgl. Bassyouny et al. (2022); Loughran und McDonald (2016, p. 1201).
[122] Vgl. Loughran und McDonald (2016, p. 1226).
[123] Vgl. Loughran und McDonald (2011, p. 49).
[124] Vgl. Henry und Leone (2016, p. 157).
[125] Vgl. Todd et al. (2024, p. 4).
[126] Vgl. Price et al. (2012, p. 1006).
[127] Vgl. Davis et al. (2014, p. 657).
[128] Vgl. Henry (2008, p. 384).
[129] Vgl. Loughran und McDonald (2016, p. 1201).
[130] Vgl. Loughran und McDonald (2016, p. 1204).

H3b: *Das Sentiment in Earnings Calls, das mit dem LM-Lexikon ermittelt wird, zeigt einen ausgeprägteren Zusammenhang zur Marktreaktion als das Sentiment, das mit den Lexika von Henry oder Harvard GI ermittelt wird*

3 Methodik zur Analyse von Earnings Calls und Marktreaktionen

3.1 Datenquellen und Beschreibung der Stichprobe

Im Rahmen der empirischen Untersuchung werden Earnings Calls von Unternehmen aus dem S&P Composite 1500 Index betrachtet, der sowohl Unternehmen des S&P 500 als auch mittelgroße und kleinere Unternehmen des S&P MidCap 400 und S&P SmallCap 600 umfasst.[131] Der Index beinhaltet insgesamt 1.506 Aktien, darunter sechs Unternehmen mit jeweils zwei Aktienklassen, und repräsentiert damit rund 90 % der Marktkapitalisierung des US-amerikanischen Aktienmarktes. Aufgrund dieser breiten Abdeckung kann er als maßgeblicher Indikator für den amerikanischen Aktienmarkt angesehen werden. Die Analyse von Unternehmen des S&P Composite 1500 ermöglicht die Untersuchung der unterschiedlichen Marktreaktionen auf Earnings Calls einer breiteren Vielfalt von Unternehmen.

Die Transkripte der Earnings Calls sowie alle weiteren Daten werden von LSEG Workspace bezogen. Um eine diversifizierte Stichprobe mit unterschiedlichen Rendite-Risiko-Profilen zu erhalten, werden für alle Unternehmen des S&P Composite 1500 Quartile für das Buchwert-zu-Markt-Verhältnis (BM-Verhältnis) und die Marktkapitalisierung gebildet. Die Marktkapitalisierung ermöglicht die Berücksichtigung von Unternehmen unterschiedlicher Größe, während das BM-Verhältnis auf mögliche Wachstumserwartungen oder Unterbewertungen hinweisen kann. Durch die Einteilung in Quartile entstehen 16 spezifische Portfolios, die Kombinationen aus Unternehmensgröße und Bewertungsverhältnis darstellen und eine breite Bandbreite von Unternehmen repräsentieren. Aus diesen wird eine Zufallsstichprobe von jeweils 15 Unternehmen gezogen. Das Portfolio 1.1, das die Unternehmen mit der niedrigsten Marktkapitalisierung und dem niedrigsten BM-Verhältnis im S&P Composite 1500 umfasst, beinhaltet jedoch lediglich zwölf Unternehmen. Daher umfasst die Gesamtstichprobe 237 anstelle der ursprünglich angenommenen 240 Unternehmen.

Der Betrachtungszeitraum erstreckt sich vom 01.01.2019 bis zum 15.03.2024. In diesem Zeitraum veröffentlichten die meisten Unternehmen ihre Earnings Calls für die Geschäftsjahre 2019 bis 2023. Da von einem Unternehmen der Stichprobe keine Earnings-Call-Transkripte vorliegen, können nur 236 Unternehmen berücksichtigt werden. Im betrachteten Zeitraum hat jedes Unternehmen durchschnittlich 20,42 Earnings Calls durchgeführt, was etwa fünf Berichtsjahren entspricht und zu einer Gesamtzahl von 4.820 analysierten Earnings Calls führt. Durchschnittlich vergingen zwischen Ende des jeweiligen Quartals und der Durchführung der Earnings Calls 34 Tage.

Die Datenverarbeitung erfolgt unter Zuhilfenahme der Statistiksoftware R sowie der darin integrierten Pakete. In einem ersten Schritt werden die Transkripte der Earnings Calls eingelesen und aufbereitet. Dabei steht die Schaffung einer einheitlichen und strukturierten Datenbasis für die nachfolgende Analyse im Vordergrund. Die Identifikation von Unternehmensnamen, Datum und Quartal erfolgt mittels spezifischer Mustererkennungstechniken unter Verwendung der Funktion „str_extract“ aus dem R-Paket „stringr“. Nach der Extraktion der relevanten Informationen werden verschiedene Datenbereinigungsprozesse durchgeführt. In diesem Zusammenhang werden unterschiedliche Bezeichnungen desselben Unternehmens in den Transkripten der verschiedenen Quartale vereinheitlicht, um eine konsistente Analyse und Kompatibilität mit anderen Daten zu gewährleisten. Darüber hinaus werden die Definitionen sowie der Disclaimer am Ende der Transkripte entfernt, da diese keinen Inhalt der ursprünglichen Telefonkonferenz wiedergeben.

Im Anschluss wird neben der Betrachtung des gesamten Earnings Calls eine Aufteilung der Transkripte in verschiedene Teile vorgenommen, um eine differenzierte Analyse des Sentiments und seiner Auswirkungen auf die Kursreaktionen zu ermöglichen. In einem ersten Schritt erfolgt eine Trennung der beiden Hauptkomponenten des Calls. Hierbei handelt es sich zum einen um die vorbereitete Präsentation und zum anderen um den anschließenden, eher spontanen Q&A-Abschnitt. Der Q&A-Abschnitt wird, wie nachfolgend beschrieben, weiter in Äußerungen von Unternehmensvertretern und Analysten untergliedert: Zu Beginn jedes Transkripts werden alle Teilnehmer mit ihren jeweiligen Funktionen aufgelistet, wobei den Äußerungen der Name des Sprechers vorangestellt ist. Um eine einheitliche Analyse zu gewährleisten, werden die Teilnehmer in zwei Kategorien eingeteilt: Unternehmensvertreter und sonstige Teilnehmer, wie Analysten. Im Rahmen der Datenaufbereitung werden die Namen im gesamten Transkript durch spezifische Platzhalter ersetzt, die den jeweiligen Teilnehmertyp kennzeichnen. Anschließend wird vor jedem dieser Platzhalter ein einheitliches Trennzeichen eingefügt, sodass die Textabschnitte effizient mithilfe der „strsplit“-Funktion segmentiert und den Funktionen der Sprecher (Manager oder Analysten) zugeordnet werden können.

Die textuelle Analyse von Earnings-Call-Transkripten erfordert eine einheitliche Struktur, um die automatische Verarbeitung, Extraktion und Zuordnung von Informationen, insbesondere in Bezug auf verschiedene Bestandteile und Personen, zu erleichtern. Von den ursprünglich 4.820 betrachteten Calls können ca. 6 % (inklusive aller Calls zweier Unternehmen) aufgrund struktureller Abweichungen nicht für die weitere Analyse verwendet werden. Nach Ausschluss der fehlerhaften Transkripte verbleiben 4.548 Earnings-Call-Transkripte, die eine konsistente und valide Datenbasis für die weiterführende Analyse bieten.

[131] Vgl. S&P Global (2024).

3.2 Sentimentermittlung in Earnings Calls

Vor der Analyse des in den Trankskripten vermittelten Sentiments ist eine Bereinigung und Vorbereitung der Texte erforderlich. In Anlehnung an die Arbeit von Allee und Deangelis (2015) erfolgt in der vorliegenden Arbeit eine spezifische Anpassung der Textabschnitte für die Analyse von Earnings Calls.[132] In diesem Zusammenhang wird das Wort „Question" entfernt, da es in Telefonkonferenzen häufig dazu verwendet wird, Bezug auf die Fragen der Analysten zu nehmen. Des Weiteren wird das Wort „good" nicht zwangsläufig als positiv kategorisiert, insbesondere wenn es in Begrüßungsformeln wie „good morning" verwendet wird, die eher konventionellen Höflichkeitsfloskeln eines Dialogs entsprechen. Ebenso wird auch der Ausdruck „Thank you" ausgeschlossen, der in vielen Fällen dazu dient, sich für die Beantwortung einer Frage zu bedanken.

In einem nächsten Schritt werden Vereinfachungsoperationen an den Textabschnitten durchgeführt, die darauf abzielen, das Datenvolumen und die Komplexität zu reduzieren, ohne dabei die Informationsdichte zu beeinträchtigen.[133] Eine Möglichkeit, einen Text zu vereinfachen, besteht darin, Stoppwörter wie Präpositionen, Artikel, Pronomen und Konjunktionen zu eliminieren.[134] Im Gegensatz zu den „lexikalischen Wörtern" (Substantive, Adjektive, Verben) enthalten diese nur wenig Information, machen jedoch etwa 20–30 % der Wörter eines Standardtextes aus. Die Umsetzung erfolgt in R mittels der Pakete „NLP" und „tm". In einem weiteren Schritt werden die zur Zuweisung der Textabschnitte zu Teilnehmerrollen verwendeten Platzhalter sowie Zahlen, Sonderzeichen und Interpunktionen entfernt. Im Anschluss erfolgt die Umwandlung aller Wörter in Kleinschreibung.

Im Rahmen einer lexikonbasierten Analyse stellt die Lemmatisierung einen weiteren wesentlichen Schritt dar.[135] Dabei werden Wörter auf ihren jeweiligen Wortstamm zurückgeführt. Dies ist insbesondere von Bedeutung, um sicherzustellen, dass alle relevanten Wörter durch das Lexikon erfasst werden, da vordefinierte Lexika häufig lediglich die Grundform der Wörter enthalten. Im Vergleich zum Stemming, bei dem das Präfix und das Suffix der Wörter entfernt werden, stellt die Lemmatisierung ein komplexeres und zeitintensiveres Verfahren dar. Sie zeichnet sich jedoch durch eine höhere Präzision sowie geringere Fehleranfälligkeit aus. Ein Beispiel für die Auswirkungen unterschiedlicher Verfahren ist die Rückführung des Wortes „better" durch Stemming auf den Wortstamm „bett". Im Gegensatz dazu identifiziert die Lemmatisierung „good" als Wortstamm. Die Funktion „lemmatize_strings" des Pakets „textstem" automatisiert die Lemmatisierung der Wörter in R. In Tabelle 1 wird exemplarisch die Bereinigung und Lemmatisierung eines Abschnitts aus einem Earnings Call dargestellt, in welchem ein Manager auf die Frage eines Analysten antwortet.

Tabelle 1: Bereinigen und lemmatisieren der Textabschnitte

Originaler Earnings Call Abschnitt	**Bereinigter und lemmatisierter Abschnitt**
Yes. We have restructured to some extent, inside the organization to focus much more on customer segments, enterprise, commercial and hyperscaler and putting specialized teams focused on driving business process to fit those 3 models in a very coordinated way across the company. And based on that, we've moved people and moved roles to different areas. And that also meant that there were some roles that were being eliminated, and that's what really caused the restructuring. And no, we don't – I mean, that was limited really to the reorganization that we put in place.[136]	*yes restructure extent inside organization focus much customer segment enterprise commercial hyperscaler put specialize team focus drive business process fit model coordinate way across company base move people move role different area also mean role eliminate really cause restructure mean limit really reorganization put place*

Im Anschluss erfolgt die Ermittlung der Anzahl positiver und negativer Wörter für jeden Textabschnitt unter Zuhilfenahme der Lexika Harvard-GI (GI), Henry (HE) sowie Loughran und McDonald (LM). Dazu wird ein Abgleich zu den Positiv-/Negativ-Listen aus den drei verschiedenen Wörterbüchern mit der Funktion „str_count" aus dem Paket „stringr" durchgeführt. Die in Tabelle 2 dargestellten Ergebnisse zeigen die Anzahl positiver (+) und negativer (-) Wörter je Wörterbuch, ermittelt aus dem exemplarischen Ausschnitt eines Earnings Calls.

[132] Vgl. Allee und Deangelis (2015, p. 10).
[133] Vgl. Tuffery (2023, p. 14).
[134] Vgl. Tuffery (2023, pp. 130f).
[135] Vgl. Tuffery (2023, pp. 124–126).
[136] Giancarlo (2024).

Tabelle 2: Quantifizierung der in den Textabschnitten enthaltenden Stimmung

Bereinigter und lemmatisierter Abschnitt	GI +	GI −	HE +	HE −	LM +	LM −
yes restructure extent inside organization focus much customer segment enterprise commercial hyperscaler put specialize team focus drive business process fit model coordinate way across company base move people move role different area also mean role eliminate really cause restructure mean limit really reorganization put place	9	9	1	0	0	2

In einem weiteren Schritt erfolgt die Transformation der Anzahl positiver ($N_{positiv}$) und negativer ($N_{negativ}$) Wörter je Earnings Call eines Unternehmens i zu dem Zeitpunkt t und Lexikon l in ein relatives Maß für das Sentiment:

$$Sentiment_{i,t,l} = \frac{N_{positiv,i,t,l} - N_{negativ,i,t,l}}{N_{positiv,i,t,l} + N_{negativ,i,t,l}}. \quad (1)$$

Die verwendete Maßeinheit für die Sentiment-Messung ist auf das Intervall [−1, 1] beschränkt. Der Wert −1 steht dabei für eine pessimistische und der Wert 1 für eine optimistische Sprache. Da die automatische Trennung verschiedener Teile des Earnings Calls (etwa in Präsentation und Q&A) mitunter zu Ungenauigkeiten führt, werden Textabschnitte, die eine Mindest- oder Höchstanzahl an Wörtern nicht erfüllen, bei der Ermittlung des Sentiments nicht berücksichtigt.

3.3 Ereignisstudie zur Untersuchung der Marktreaktionen

Die Frequenz, in der Unternehmen Berichte veröffentlichen – typischerweise in Quartal- oder Jahresschritten – begrenzt die Eignung von aus Unternehmensveröffentlichungen ermitteltem Sentiment als Input für Zeitreihenmodelle.[137] Dennoch eignen sich diese Daten für Querschnittsanalysen und Ereignisstudien, in denen die Auswirkungen der Stimmung auf die Kapitalmarktreaktion rund um Unternehmensveröffentlichungen untersucht werden können. Die Ereignisstudien-Methodik, wie sie heute in der akademischen Forschung Anwendung findet, wurde erstmals 1969 von Fama et al. als Instrument zur Ermittlung ereignisinduzierter Renditen eingeführt.[138] Sie erlaubt die Quantifizierung des Einflusses eines Ereignisses auf die Rendite einer Aktie innerhalb eines definierten Zeitraums, dem sogenannten „Ereignisfenster“. Folglich kann mit ihrer Hilfe die Marktbewertung des Informationsgehalts von Offenlegungen sowie der allgemeinen Qualität der Finanzberichterstattung erfasst und analysiert werden.[139]

Ereignisstudien basieren auf der Prämisse einer mittelstarken Markteffizienz, bei der Investoren auf die Veröffentlichung oder das Bekanntwerden neuer Informationen (das Ereignis) reagieren.[140] Eine durch ein Ereignis ausgelöste Änderung des Aktienkurses erfolgt, wenn die zugrunde liegende Information einen positiven oder negativen Informationswert besitzt und für die Marktteilnehmer neu und unerwartet ist.[141] Im Vergleich zu älteren Methoden berücksichtigt die moderne Ereignisstudien-Methodik nicht einen Vorher-Nachher-Kursvergleich, sondern die hypothetische Kursentwicklung des Unternehmens, falls das Ereignis nicht eingetreten wäre.[142] Die Differenz zwischen der beobachteten Kursveränderung $R_{i,t}$ und der hypothetisch erwarteten Rendite $E(R_{i,t})$ wird als abnormale Rendite $AR_{i,t}$ bezeichnet und dient als Indikator für die Kapitalmarktbewertung eines Ereignisses:

$$AR_{i,t} = R_{i,t} - E(R_{i,t}). \quad (2)$$

In der wissenschaftlichen Literatur finden sich zahlreiche Methoden zur Bestimmung der erwarteten Rendite $E(R_{i,t})$.[143] Die resultierenden Referenzrenditen können dabei erheblich variieren. Das von Sharpe (1963) auf Basis der Markowitz-Portfolio-Theorie entwickelte Marktmodell stellt eine Variante des Einfaktormodells dar und gilt als nahezu universell einsetzbares Instrument zur Bestimmung erwarteter Renditen in kurzfristigen Eventstudien.[144] Dieses Modell erweist sich insbesondere dann als geeignet, wenn das untersuchte Ereignis, wie auch Earnings Calls, in erster Linie unternehmensspezifisch ist und keine wesentlichen Auswirkungen auf den Gesamtmarkt oder systematische Risiken aufweist.[145] Die Marktrendite $R_{m,t}$ stellt dabei die wesentliche Einflussgröße auf die Rendite des untersuchten Wertpapiers dar.[146] Im Rahmen dieser Arbeit wird die Rendite des S&P Composite 1500 als relevante Marktrendite $R_{m,t}$ definiert. Die Formel zur Bestimmung der erwarteten Rendite lautet:

$$E(R_{i,t}) = \widehat{\alpha}_i + \widehat{\beta}_i * R_{m,t}, \quad (3)$$

[137] Vgl. Kearney und Liu (2014, p. 174).
[138] Vgl. Holler (2016, p. 11).
[139] Vgl. Henry (2008, p. 283).
[140] Vgl. Gehrke (2022, p. 279).
[141] Vgl. Holler (2016, p. 23).
[142] Vgl. Gerpott (2009, p. 206).
[143] Vgl. Gehrke (2022, p. 279).
[144] Vgl. Henry und Leone (2016, p. 277); Holler (2016, pp. 82f).
[145] Vgl. Müller (2023).
[146] Vgl. Gehrke (2022, p. 280).

wobei die Koeffizienten $\widehat{\alpha}_i$ und $\widehat{\beta}_i$ eines Unternehmens i für jeden Zeitpunkt t durch eine lineare Regression aus historischen Unternehmens- und Marktrenditen ermittelt werden. Um eine Verzerrung durch das Ereignis selbst zu vermeiden, muss die Schätzung für die Regression des Markmodells dabei außerhalb des Ereigniszeitraums liegen. In der vorliegenden Untersuchung wird die Schätzperiode von 202 bis 2 Handelstage vor dem Earnings Call für die Schätzung der normalen Renditen herangezogen. Die Spanne von 200 Handelstagen entspricht hierbei der gängigen Praxis in der Ereignisstudienliteratur.[147]

Im Rahmen der vorliegenden Arbeit werden logarithmierte Renditen berücksichtigt. Diese werden gemäß der Formel:

$$R_{i,t} = \ln\left(\frac{Schlusskurs_{i,t}}{Schlusskurs_{i,t-1}}\right) \tag{4}$$

berechnet, wobei $Schlusskurs_{i,t}$ den Schlusskurs der Aktie am Tag t und $Schlusskurs_{i,t-1}$ den einen Handelstag zuvor darstellt. Diese Methode bietet im Vergleich zu diskreten Renditen mehrere Vorteile: Logarithmierte Renditen sind mathematisch einfacher zu handhaben, zeigen bessere statistische Eigenschaften und neigen stärker zur Annäherung an eine Normalverteilung.[148] Aufgrund dieser Vorteile wird auch für Eventstudien häufig die Verwendung logarithmierter Renditen empfohlen.

Die kumulierte abnormale Rendite (CAR) bezeichnet die aufsummierte abnormale Rendite $AR_{i,t}$ eines Unternehmens i innerhalb eines definierten Ereignisfensters.[149] Die allgemeine Formel lautet:

$$CAR(t_1, t_2)_i = \sum_{t=t_1}^{t_2} AR_{i,t}, \tag{5}$$

wobei t=0 der Tag des Earnings Calls ist, t_1 die untere Grenze und t_2 die obere Grenze des Ereignisfensters bezeichnet. Für die vorliegende Untersuchung wird mit CAR(-1,1) das initiale Ereignisfenster auf drei Tage um den Termin des Calls herum festgelegt. In Anlehnung an frühere Erkenntnisse, denen zufolge der Markt am Tag t-1 signifikant auf Unternehmensveröffentlichungen reagiert, beginnt das Ereignisfenster am Tag vor dem Call.[150] Die Verwendung eines dreitägigen Ereignisfensters bis einen Tag nach der Telefonkonferenz ist konsistent mit früheren Studien zur initialen Marktreaktion auf Informationen in Earnings Calls.[151] Das Ereignisfenster umfasst anstelle von Kalendertagen Börsentage. Dies dient der Gewährleistung einer einheitlichen Definition des Zeitraums, innerhalb dessen der Kapitalmarkt in der Lage ist, neue Informationen zu verarbeiten.[152]

Des Weiteren wird neben dem initialen Reaktionsfenster auch die nachfolgende Driftperiode in die Betrachtung einbezogen. Ball und Brown (1968) zeigten, dass die CAR nach Gewinnankündigungen bei als „gut“ klassifizierten Nachrichten im Zeitverlauf weiter ansteigen, während sie bei als „schlecht“ bewerteten Nachrichten weiter sinken.[153] Dieses Phänomen wird als Drift nach der Gewinnankündigung (Post-Earnings Announcement Drift, PEAD) bezeichnet. Die Resultate langfristiger Ereignisstudien legen die Vermutung nahe, dass die anfänglichen Marktreaktionen veröffentlichte Unternehmensinformationen nicht vollständig widerspiegeln.[154] Die Driftperiode auf den Kapitalmärkten ist der Zeitraum nach einer Ereignisankündigung, in dem Kursbewegungen weiter anhalten oder sich gegenläufig entwickeln können. Während die Märkte anfänglich schnell auf neue Informationen reagieren, verändern sich in der anschließenden Driftperiode die Auswirkungen des Ereignisses durch fortlaufende Informationsverarbeitung und Marktpsychologie. Diese Phase ist entscheidend, um die Effektivität der Informationsverarbeitung durch den Markt sowie die zeitliche Entwicklung der Einschätzungen der Investoren zu verstehen. Zu diesem Zweck werden in dieser Arbeit die kumulierten abnormalen Renditen CAR(2,42) zwischen dem zweiten und 42. Handelstag nach den Earnings Calls analysiert.

3.4 Einflussfaktoren auf Aktienrenditen während Earnings Calls

Für vier Unternehmen in der Stichprobe konnten keine Kontrollvariablen ermittelt werden, sodass insgesamt 230 Unternehmen für die weitere Analyse verbleiben. Des Weiteren liegen teilweise für die verbleibenden Unternehmen keine Rendite- oder Kontrolldaten in Bezug auf einzelne Earnings Calls vor, sodass die weitere Analyse auf 4.190 Beobachtungen basiert.

Die grundlegende Struktur dieser Untersuchung basiert auf einer Regressionsanalyse, bei der CAR um die Zeit des Earnings Calls als abhängige Variable definiert werden. Die zentrale unabhängige Variable ist das mit dem Lexikon l ermittelte $Sentiment_{i,q,l}$ in dem Earnings Call von Unternehmen i zum Quartal q:

$$\begin{aligned} CAR(t_1, t_2)_{i,q} =& \alpha + \beta * Sentiment_{i,q,l} + \sum Controls_{i,q} \\ &+ \delta_i + \lambda_q + \varepsilon_{i,q}. \end{aligned} \tag{6}$$

Zur Evaluierung der Reaktion von Investoren hinsichtlich des Tons eines Konferenzgesprächs erfolgt eine Schätzung von Panelregressionen mit festen Effekten für Unternehmen δ_i und Quartale λ_q. Des Weiteren wird der Ansatz der robusten Standardfehler nach White verfolgt, um potenzielle Heteroskedastizität in den Daten adäquat zu berücksichtigen.

[147] Vgl. Holler (2016, pp. 38f).
[148] Vgl. Gehrke (2022, p. 279); Holler (2016, p. 80).
[149] Vgl. Gehrke (2022, pp. 283f).
[150] Vgl. Henry (2008, p. 358).
[151] Vgl. Davis et al. (2014, p. 669); Price et al. (2012, p. 998).
[152] Vgl. Gerpott (2009, p. 214).
[153] Vgl. Ball und Brown (1968, p. 170).
[154] Vgl. H. K. Baker und Nofsinger (2010, p. 338).

Auf die Berücksichtigung robuster Standardfehler für Autokorrelation wird verzichtet, da der Wooldridge-Test keinen Hinweis auf serielle Korrelation in den Residuen liefert.

Der Vektor $\sum Controls_{i,q}$ stellt eine Sammlung von Variablen auf Unternehmensebene zum Ende des jeweiligen Quartals q dar, auf das sich der Earnings Call bezieht. Durch die Aufnahme von Kontrollvariablen wird die Wahrscheinlichkeit verringert, dass die auf Wörterbüchern basierenden Sentiment-Maße Informationen erfassen, die mit anderen beobachtbaren Unternehmensmerkmalen am Datum der Telefonkonferenz zusammenhängen. Die Gewinnüberraschung bezeichnet die Differenz zwischen dem ausgewiesenen Gewinn eines Unternehmens und den Erwartungen des Marktes.[155] Sie stellt einen der maßgeblichen Einflussfaktoren bei der Bekanntgabe von Unternehmensergebnissen dar. Im Rahmen der vorliegenden Arbeit erfolgt die Ermittlung der Gewinnüberraschung mittels einer saisonalen Messung. Die Gewinnüberraschung eines Unternehmens i im Quartal q lässt sich demnach wie folgt berechnen:

$$SURP_{i,q} = \frac{(EPS_{i,q} - EPS_{i,q-4})}{Schlusskurs_{i,q-4}}, \tag{7}$$

wobei $EPS_{i,q}$ den Gewinn je Aktie des aktuellen Quartals, $EPS_{i,q-4}$ den Gewinn je Aktie des entsprechenden Vorjahresquartals und $Schlusskurs_{i,q-4}$ den Schlusskurs der Aktie zum Ende des betreffenden Quartals bezeichnet. Diese Vorgehensweise wird gegenüber einer auf Analystenprognosen basierenden Ermittlung bevorzugt, da der Gewinn im gleichen Quartal des Vorjahres aus der Perspektive des Managements eine bedeutsamere Benchmark darstellt als die durchschnittlichen Schätzungen der Analysten und letztere in der Regel Gewinne prognostizieren, die nicht im Einklang mit standardisierten Rechnungslegungsgrundsätzen stehen.[156] Die Verwendung einer saisonalen Messung der Gewinnüberraschung steht darüber hinaus im Einklang mit anderen Studien, die das Sentiment in Earnings Calls analysieren, und ermöglicht eine konsistente Vergleichbarkeit der Ergebnisse.[157]

SIZE bezieht sich auf die Unternehmensgröße, die anhand der logarithmierten Marktkapitalisierung in Millionen US-Dollar gemessen wird. Mit steigender Unternehmensgröße nimmt die Menge an verfügbaren Informationen zu, wobei gleichzeitig eine potenziell stärkere regulatorische und öffentliche Kontrolle besteht.[158] Das BM-Verhältnis (BM) misst das Verhältnis des Buchwerts des Eigenkapitals zu der Marktkapitalisierung eines Unternehmens. Ein höheres Verhältnis deutet auf eine Unterbewertung des Unternehmens hin, während ein niedrigeres Verhältnis auf ein wachstumsorientiertes Unternehmen hinweist.[159] Mit dem Return on Assets (ROA) lässt sich die Effektivität eines Unternehmens bei der Nutzung seiner Vermögenswerte zur Gewinnmaximierung beurteilen, was besonders bei nicht profitablen Unternehmen von Bedeutung ist. Ermittelt wird er durch Division des Nettogewinns durch die Gesamtsumme der Aktiva und anschließende Multiplikation des Ergebnisses mit 100. Die Kennzahl LEVERAGE gibt das Verhältnis von Gesamtschulden zu den jeweiligen Gesamtvermögenswerten eines Unternehmens an. Ein hoher Wert indiziert ein erhöhtes Insolvenzrisiko, während eine moderate Verschuldung auf eine effektive Nutzung des Fremdkapitals für Unternehmenswachstum hinweist.[160] LOSS ist eine Dummy-Variable, die den Wert 1 annimmt, sofern das Unternehmen im betrachteten Quartal einen Verlust erwirtschaftet hat, und den Wert 0 in allen anderen Fällen. Unternehmen, die Verluste verzeichnen, werden in der Regel als riskanter eingestuft, was sich in einer negativen Rendite niederschlagen kann.[161] GROWTH bezeichnet das prozentuale Umsatzwachstum eines Unternehmens im Vergleich zum vorherigen Quartal. Die Variable wird zur Glättung von Extremwerten sowie zur Erzielung einer symmetrischeren Verteilung als natürlicher Logarithmus von eins plus Umsatzwachstumsrate berechnet. Ein signifikantes Umsatzwachstum kann als Indikator für eine positive Geschäftsentwicklung sowie vielversprechende Zukunftsaussichten gewertet werden.[162] Die Jahre seit dem Börsengang werden durch FIRMAGE erfasst, berechnet als natürlicher Logarithmus von eins plus der Anzahl der Jahre seit dem Börsengang. In den ersten Jahren nach dem Börsengang eines Unternehmens sind historische Finanzdaten und Informationen für die Öffentlichkeit oft nicht verfügbar.[163] Dies resultiert in einer höheren Informationsasymmetrie sowie einem gesteigerten Risiko. Die Anzahl an Analysten, die an einer Konferenz teilnehmen, wird von ANALYSTS angegeben. Eine hohe Beteiligung von Analysten lässt auf ein großes Interesse am Unternehmen und dessen Bedeutung für die Finanzmärkte schließen.[164] Zudem lässt sich beobachten, dass die Teilnahme von Analysten den Informationsfluss und die Qualität, der in den Earnings Calls vermittelten Informationen beeinflusst.

4 Ergebnisse der Ereignisstudie zum Sentiment in Earnings Calls

4.1 Sentiment in Earnings Calls

4.1.1 Deskriptive Statistiken des Sentiments in Earnings Calls

Die Gegenüberstellung der Median- und Mittelwerte der drei Sentimentvariablen zeigt Unterschiede in der Stimmungsermittlung zwischen den drei verschiedenen Lexika. Das LM-Lexikon weist mit einem Medianwert von

[155] Vgl. Price et al. (2012, p. 998).
[156] Vgl. Price et al. (2012, p. 998).
[157] Vgl. Milian et al. (2017, p. 70); Blau et al. (2015, p. 208); Brockman et al. (2015, p. 26); Price et al. (2012, p. 998).
[158] Vgl. Bochkay et al. (2019, p. 49).
[159] Vgl. Price et al. (2012, p. 999).
[160] Vgl. Price et al. (2012, p. 999).
[161] Vgl. Mather et al. (2021); Davis et al. (2014, p. 657).
[162] Vgl. Yamamoto et al. (2022, p. 249); Bochkay et al. (2019, p. 53); Blau et al. (2015, p. 214); Davis et al. (2014, p. 657).
[163] Vgl. Ning et al. (2023).
[164] Vgl. Cicon (2015, p. 72).

-0,0137 auf ein insgesamt leicht negatives Sentiment in den Earnings Calls hin, was auf eine tendenziell pessimistische Grundstimmung schließen lässt. Im Gegensatz dazu zeigen das Harvard-GI-Lexikon und das Henry-Lexikon mit Medianwerten von 0,09561 bzw. 0,5875 eine überwiegend positive Stimmung. Die durchschnittlichen monatlichen Sentiment-Werte je Lexikon im Betrachtungszeitraum sind in Abbildung 1 dargestellt und veranschaulichen die Unterschiede zwischen den Lexika.

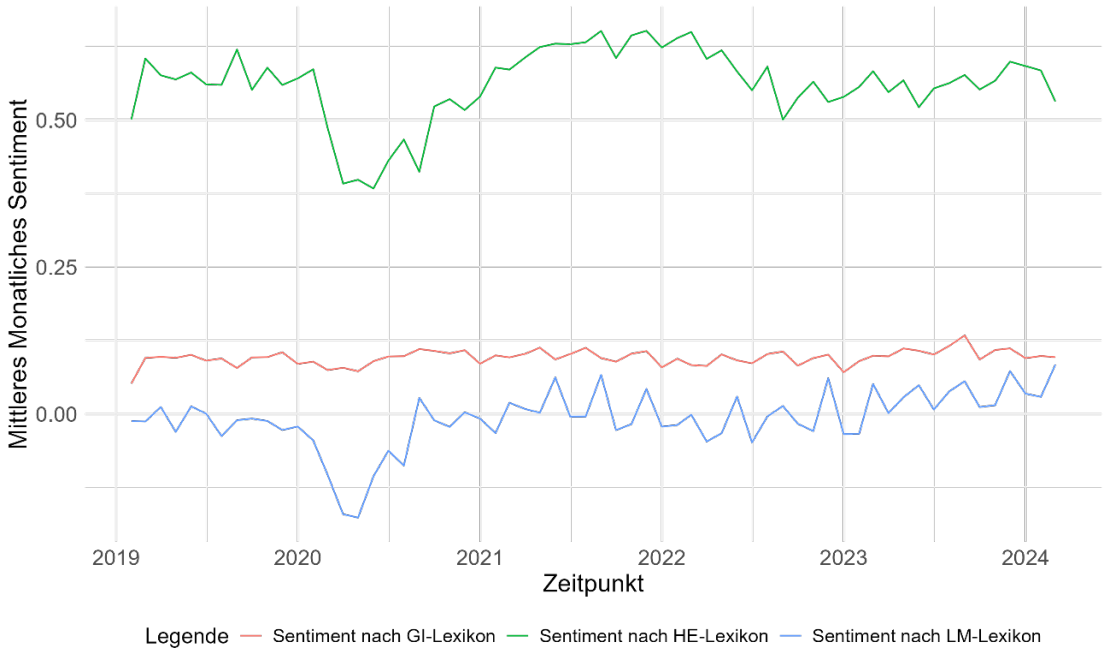

Abbildung 1: Mittleres Monatliches Sentiment in Earnings Calls nach Loughran McDonald (LM), Henry (HE) und Harvard-GI (GI) Lexikon

Die Betrachtung der Spannweite der Sentiment-Werte verdeutlicht, dass die Stimmung in den Earnings Calls sowohl starke negative als auch starke positive Ausprägungen annimmt. Der niedrigste gemessene Wert aller Lexika von -0,6238 wurde mit dem LM-Lexikon für einen Earnings Call des ersten Quartals 2022 von Medpace Holdings ermittelt. Obwohl die Gewinnerwartungen leicht übertroffen wurden, fokussierte sich die Diskussion in der Konferenz primär auf Verzögerungen und Stornierungen von Programmen, was Besorgnis hinsichtlich der künftigen Wachstumspläne auslöste. Im Gegensatz dazu wurde der höchste Sentiment-Wert von 0,9080 für einen Earnings Call des ersten Quartals 2021 von West Pharmaceutical Services Inc. mit dem Henry-Lexikon erfasst. Im Rahmen der Veranstaltung wurden Informationen zu einer hohen Nachfrage nach den Produkten in Verbindung mit der Coronavirus-Pandemie diskutiert.

Die Unterschiede in den deskriptiven Statistiken lassen sich in großen Teilen auf die unterschiedlichen Verhältnisse von positiven zu negativen Begriffen in den Lexika zurückführen. Im LM-Lexikon liegt dieses Verhältnis bei 1:6,65, was bedeutet, dass es wesentlich mehr negative als positive Wörter umfasst. Dies führt zu den niedrigsten und negativsten gemessenen Sentiment-Werten unter den untersuchten Lexika. Demgegenüber zeigen das GI- und das HE-Lexikon ein ausgewogeneres bzw. umgekehrtes Verhältnis von positiven zu negativen Begriffen. Im GI-Lexikon ist die Anzahl der negativen Begriffe leicht höher als die Anzahl der positiven Begriffe (Verhältnis 1:1,22). Im Gegensatz dazu zeigt sich im HE-Lexikon ein leichtes Überwiegen positiver Begriffe (Verhältnis 1,24:1). Diese unterschiedlichen Verhältnisse manifestieren sich in einer positiveren Tendenz hinsichtlich der gemessenen Sentiment-Werte. Darüber hinaus werden die geringsten Standardabweichungen im GI-Lexikon beobachtet, was auf eine konsistentere Bewertung des Sentiments hinweist. Demgegenüber weisen das LM- und das Henry-Lexikon höhere Abweichungen auf, was auf eine differenziertere Analyse zurückzuführen sein könnte. Die Varianz der Sentiment-Werte in Abhängigkeit vom verwendeten Lexikon verdeutlicht die Bedeutung der spezifischen Eigenschaften sowie des Gewichtungsverhältnisses der Lexika bei der Analyse der Beziehung zwischen Sentiment und Marktreaktionen. Die deskriptiven Statistiken legen insgesamt nahe, dass die Auswahl des Lexikons einen signifikanten Einfluss auf die gemessenen Sentiment-Werte hat, was bei der weiteren Analyse der abnormalen Renditen zu berücksichtigen ist.

In Anbetracht der unterschiedlichen Inhalte und Sprecherperspektiven ist es entscheidend, das jeweilige Sentiment des Earnings Calls differenziert zu betrachten. Hierbei ist insbesondere eine Unterscheidung zwischen der

allgemeinen Stimmung während des Calls und den spezifischen Segmenten wie Präsentation und Q&A oder Rollen wie Manager und Analysten erforderlich. Im Henry-Lexikon wird die Präsentation tendenziell positiver bewertet als das Q&A (Median: 0,5966 vs. 0,5789). In diesem Lexikon fällt der Unterschied zwischen den beiden Abschnitten jedoch weniger deutlich aus als in den anderen untersuchten Lexika. Der im Vergleich zur Präsentation negativere Q&A-Abschnitt ist konsistent mit ähnlichen Befunden in der Literatur, wonach diese Beobachtung nicht auf das Henry-Lexikon beschränkt ist.[165] Das Beantworten kritischer und detaillierter Fragen während der Q&A-Sitzung kann demzufolge dazu führen, dass das Management zu spontanen und ehrlichen Antworten gezwungen wird, die weniger optimistisch ausfallen als ihre vorbereiteten Präsentationen. Entgegen dieser Erwartung weisen das Harvard-GI- und das LM-Lexikon in dieser Arbeit eine positivere Tendenz im Frage-Antwort-Abschnitt (Median LM: 0,0698, GI: 0,1039) als in der Präsentation (Median LM: -0,0988, GI: 0,0887) auf. Dies deutet darauf hin, dass die Fragen der Analysten selbst eine positivere Formulierung aufweisen oder dass die Unternehmensführung während der Q&A-Sitzung optimistischere oder beruhigendere Antworten als in der Präsentation gibt.

Tabelle 3: Deskriptive Statistiken

	Variabel	Mittelwert	Median	Min.	Max.	SD
Sentiment	*Call (LM)*	−0.0190	−0.0137	−0.6238	*0.5190*	*0.1511*
	Call (HE)	*0.5644*	*0.5875*	−0.1341	*0.9080*	*0.1396*
	Call (GI)	*0.0967*	*0.0956*	−0.0821	*0.3461*	*0.0558*
	Präsentation (LM)	−0.1081	−0.0988	−0.8272	*0.4648*	*0.1822*
	Präsentation (HE)	*0.5654*	*0.5966*	−0.1833	*0.9192*	*0.1751*
	Präsentation (GI)	*0.0900*	*0.0887*	−0.1463	*0.3702*	*0.0666*
	Q&A (LM)	*0.0675*	*0.0698*	−0.6403	*0.6059*	*0.1748*
	Q&A (HE)	*0.5642*	*0.5789*	−0.1023	*1.0000*	*0.1465*
	Q&A (GI)	*0.1044*	*0.1039*	−0.1137	*0.3787*	*0.0676*
	Manager Q&A (LM)	*0.1024*	*0.1060*	−0.6552	*0.6897*	*0.1850*
	Manager Q&A (HE)	*0.6171*	*0.6309*	−0.0909	*1.0000*	*0.1513*
	Manager Q&A (GI)	*0.1043*	*0.1053*	−0.1644	*0.4118*	*0.0787*
	Analyst Q&A (LM)	−0.0108	*0.0000*	−1.0000	*1.0000*	*0.2691*
	Analyst Q&A (HE)	*0.4094*	*0.4254*	−0.6667	*1.0000*	*0.2237*
	Analyst Q&A (GI)	*0.0943*	*0.0934*	−0.3247	*0.5023*	*0.0881*
Kontrollvariabeln	*SURP*	*0.0044*	*0.0010*	−0.4215	*1.4319*	*0.0512*
	SIZE	*8.7316*	*8.5499*	*4.8819*	*14.8433*	*1.5387*
	BM	*0.4394*	*0.3318*	−2.7550	*8.7767*	*0.5640*
	ROA	*0.0518*	*0.0504*	−0.6283	*0.5078*	*0.0775*
	LEVERAGE	*0.2863*	*0.2764*	*0.0000*	*2.2923*	*0.1963*
	GROWTH	*0.0164*	*0.0183*	−4.1044	*3.6384*	*0.2354*
	LOSS	*0.1649*	*0.0000*	*0.0000*	*1.0000*	*0.3711*
	FIRMAGE	*3.1094*	*3.2594*	*0.7396*	*4.6509*	*0.7854*
CAR	*ANALYSTS*	*7.6737*	*7.0000*	*1.0000*	*24.0000*	*3.7025*
	CAR(-1,1)	*0.0010*	*0.0042*	−0.4587	*0.5630*	*0.0869*
	CAR(2,42)	*−0.0060*	*−0.0081*	−1.2156	*1.2957*	*0.1446*

Innerhalb des Frage-und-Antwort-Abschnitts weisen alle drei Lexika einen positiveren Trend in den Sentiment-Werten der Aussagen von Managern in dem Q&A im Vergleich zu den Aussagen von Analysten auf. Die Medianwerte des Manager-Sentiments sind höher als die der Analysten-Sentiments. Insbesondere im Henry-Lexikon zeigt sich eine ausgeprägte Differenz, während der Unterschied im GI-Lexikon weniger deutlich ist. Diese Beobachtung differierender Haltungen zwischen Managern und Analysten deckt sich mit den Resultaten vorheriger Studien.[166] Die hohe Streuung der Sentiment-Werte der Analysten im Q&A-Teil deutet darauf hin, dass Analysten während des Q&A-Teils stark unterschiedliche Einschätzungen und Reaktionen aufweisen. Dabei legt die positivere Tonalität der Manager im Q&A-Abschnitt im Vergleich zu den Analysten nahe, dass Unternehmensvertreter versuchen, negative Fragestellungen durch eine optimistische Argumentation zu kompensieren. Des Weiteren lässt sich eine positivere Stimmung in den

[165] Vgl. De Amicis et al. (2021); Fu et al. (2019, p. 649); Borochin et al. (2018, p. 81); Brockman et al. (2015, p. 27)

[166] Vgl. Fu et al. (2019, p. 649); Borochin et al. (2018, p. 81); Brockman et al. (2015, p. 27)

Aussagen der Manager in den Q&A-Abschnitten aller drei Lexika im Vergleich zu ihren Aussagen in der Präsentation feststellen. Dies deutet darauf hin, dass Manager während tendenziell offenherziger und positiver auf Fragen eingehen als in der formalen Präsentation, was zu einem insgesamt höheren Sentiment-Wert in diesem Abschnitt des Calls führt.

Die deskriptiven Statistiken für alle abhängigen, erklärenden und Kontrollvariablen sind in Tabelle 3 dargestellt.

Die anfänglichen Marktreaktionen CAR(-1,1) um den Tag des Earnings Calls variieren zwischen -0,4587 und 0,5630. Diese Spannweite lässt auf Unterschiede in der Relevanz der veröffentlichten Informationen sowie divergierende Erwartungen der Investoren schließen. Der ermittelte Median von 0,0042 deutet auf tendenziell positive Marktreaktionen hin. Im Gegensatz dazu lässt die Driftperiode CAR(2,42) nach den Earnings Calls mit einem Median von -0,0081 eine leicht negative Tendenz erkennen. Die signifikante Spannweite demonstriert, dass einige Unternehmen in dieser Periode erhebliche Verluste oder Gewinne verzeichnen. Ein Vergleich der Werte von CAR(1,1) mit denen von CAR(2,42) zeigt, dass die letztgenannten Werte etwa doppelt so hoch sind. Dies lässt den Schluss zu, dass der Informationsgehalt der Earnings Calls während der anfänglichen Reaktionszeit nicht vollständig in den Kursen enthalten ist. Die tendenziell positiven direkten Reaktionen auf Earnings Calls und die oft negative nachfolgende Performance weisen auf eine komplexe Dynamik zwischen kurzfristigen Marktreaktionen und längerfristigen Entwicklungen hin.

Die Analyse der Kontrollvariablen zeigt, dass der Median der Gewinnüberraschungen bei 0,0010 liegt, was auf eine leichte Tendenz zu positiven Überraschungen hindeutet, die mit früheren Studien übereinstimmt.[167] Das BM-Verhältnis beträgt 0,4394, der Return on Assets 0,0518 und der Verschuldungsgrad 0,2863. Die durchschnittliche logarithmierte Marktkapitalisierung der untersuchten Unternehmen beträgt 8,7316 Millionen US-Dollar. Das Umsatzwachstum variiert zwischen -4,1043 und 3,6384, was eine breite Spanne an Wachstumsraten impliziert. Das durchschnittliche Alter der betrachteten Unternehmen, berechnet als natürlicher Logarithmus von 1 plus der Jahre seit dem Börsengang, beträgt 3,1094. Die Mehrheit der Unternehmen ist folglich bereits seit mehreren Jahren börsennotiert, während einige ihren Börsengang während des Beobachtungszeitraums durchgeführt haben. Die Anzahl der an den Calls teilnehmenden Analysten variiert zwischen einem und 24, wobei ein Durchschnitt von etwa acht Analysten pro Call zu verzeichnen ist.

4.1.2 Korrelationen des Sentiments in Earnings Calls

Wie der Korrelationsmatrix in Anhang 1 zu entnehmen sind die Korrelationen zwischen den Tonmaßen der verschiedenen Lexika durchweg positiv, jedoch nicht perfekt. Dies lässt den Schluss zu, dass die Lexika zum einen teilweise ähnliche, zum anderen aber auch unterschiedliche Informationen zur Stimmung erfassen. Besonders hohe Übereinstimmungen zeigen sich zwischen dem LM- und dem Henry-Lexikon, insbesondere bei der Erfassung des Gesamt-Sentiments im Call und während der Präsentation. Des Weiteren ist eine Varianz der Lexika in Bezug auf das Korrelationsniveau zwischen den ermittelten Tonmaßen und den kumulierten abnormalen Renditen im initialen Reaktionsfenster sowie der Driftperiode festzustellen. Die Korrelationen der Lexika mit CAR(-1,1) sind durchweg positiv. Das Henry-Lexikon weist hierbei die höchsten Werte auf. Eine Ausnahme bildet die höchste Gesamtkorrelation zwischen dem Sentiment der Analysten im Q&A, ermittelt mit dem LM-Lexikon, und CAR(-1,1). In der Driftperiode sind die Korrelationen des Sentiments mit CAR(2,42) insgesamt geringer und teilweise negativ.

Die Analyse der entsprechenden Tonmaße ergibt, dass in den verschiedenen Abschnitten und bei den unterschiedlichen Sprechern innerhalb der Earnings Calls sprachliche Nuancen zu beobachten sind. Die positiven Korrelationen zwischen den Tonvariablen in verschiedenen Abschnitten (Präsentation, Q&A) und Rollen der Redner (Unternehmensvertreter, Analysten) bei allen drei Lexika lassen den Schluss zu, dass eine gemeinsame Signalkomponente existiert, die die übergeordnete Stimmung während der gesamten Telefonkonferenz reflektiert. Diese gemeinsame Komponente ist jedoch nicht perfekt, da die Korrelationen zwischen den verschiedenen Abschnitten und Rednerrollen mit gleichem Lexikon zwar positiv sind, aber zwischen 0,2219 und 0,5107 variieren. Die geringsten Korrelationen sind zwischen dem Sentiment in der Präsentation und dem Sentiment der Analysten zu beobachten. Dies lässt den Schluss zu, dass Analysten und Management unterschiedliche Perspektiven und divergierende Erwartungen haben.

Das Sentiment im Q&A weist, mit Ausnahme der Analyse mit dem GI-Lexikon, im initialen Reaktionsfenster eine stärkere positive Korrelation mit CAR(-1,1) auf als das Sentiment in der Präsentation. Im Vergleich dazu lässt sich in der Driftperiode CAR(2,42) eine leicht stärkere negative Korrelation des Sentiments der Präsentation mit den abnormalen Renditen beobachten, was auf eine unterschiedlich starke Wirkung der beiden Abschnitte des Earnings Calls hindeutet. Eine Gegenüberstellung des Sentiments innerhalb des Q&A zeigt, dass das Sentiment der Analysten im initialen Reaktionsfenster eine stärkere Korrelation mit den CAR(-1,1) aufweist als das Sentiment der Manager. Das Analysten-Sentiment erzielt sowohl für das LM- als auch für das Henry-Lexikon die höchsten Korrelationswerte zu CAR unter allen untersuchten Variablen. In der Driftperiode CAR(2,42) lassen sich keine signifikanten Unterschiede zwischen den Einschätzungen der Analysten und der Manager im Q&A feststellen.

Die Korrelationen zwischen den Kontrollvariablen und den abnormalen Renditen sind insgesamt von geringer Intensität, was darauf hindeutet, dass diese Variablen nur eine untergeordnete Rolle bei der Erklärung der abnormalen Renditen spielen. Darüber hinaus lässt die geringe Korrelation zwischen Sentiment und Kontrollvariablen den Schluss zu, dass das Sentiment unabhängige Informationen liefert.

[167] Vgl. Milian et al. (2017, p. 70); Blau et al. (2015, p. 209); Brockman et al. (2015, p. 27); Price et al. (2012, p. 1000).

In Anbetracht der Größenordnungen der meisten Korrelationen ist davon auszugehen, dass Multikollinearität bei der Auswertung der nachfolgenden Regressionen keine wesentlichen Schwierigkeiten bereiten wird. Eine Betrachtung der Korrelationsmatrix zeigt, dass keine starken Korrelationen zwischen den unabhängigen Variablen beobachtet werden. Lediglich zwischen Variablen, die nicht gleichzeitig in den Regressionen berücksichtigt werden, treten starke Korrelationen von über 0,7 auf. Dies betrifft beispielsweise das Sentiment des gesamten Calls und das der Präsentation. Die stärksten Korrelationen zwischen gleichzeitig betrachteten Variablen bestehen zwischen den Kontrollvariablen ROA und LOSS mit einem Wert von -0,53 sowie zwischen der Unternehmensgröße und der Anzahl der teilnehmenden Analysten.

Tabelle 4: Einfluss des Gesamtsentiments in Earnings Calls sowie des differenzierten Sentiments in Präsentation und Q&A auf CAR(-1,1)

	LM	HE	GI	LM	HE	GI
Gesamter Call	*0.136****	*0.148****	*0.281****			
	(0.0144)	*(0.0157)*	*(0.0439)*			
Präsentation				*0.054****	*0.066****	*0.153****
				(0.0141)	*(0.0134)*	*(0.0341)*
Q&A				*0.067****	*0.073****	*0.103****
				(0.0128)	*(0.0141)*	*(0.0309)*
R^2	*0.0737*	*0.0809*	*0.0630*	*0.0812*	*0.0888*	*0.0722*
Adjusted R^2	*0.0116*	*0.0193*	*0.0002*	*0.0132*	*0.0214*	*0.0035*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

4.2 Einfluss des Sentiments in Earnings Calls auf kumulierte abnormale Renditen

4.2.1 Gesamtsentiment

Im Rahmen dieser Untersuchung wurden Panelregressionen durchgeführt, um den Zusammenhang zwischen kumulierten abnormalen Renditen im initialen Ereignisfenster CAR(-1,1) und dem Sentiment in Earnings Calls zu analysieren. Das Sentiment wurde jeweils anhand der Lexika LM, Henry und Harvard-GI in einem ersten Schritt für den gesamten Earnings Call ermittelt:

$$CAR(-1,1)_{i,q} = \alpha + \beta * Sentiment_{Gesamt,i,q,l} + \sum Controls_{i,q} + \delta_i + \lambda_q + \varepsilon_{i,q}. \quad (8)$$

Die in Tabelle 4 dargestellten Ergebnisse aller Modelle legen nahe, dass ein positiveres Sentiment einen signifikant positiven Einfluss (auf 0,1-%-Niveau) auf die CAR(-1,1) ausübt. Auf dieser Basis lässt sich die Hypothese H1a bestätigen, dass ein positives Sentiment in Earnings Calls signifikant positive Marktreaktionen hervorruft, während ein negatives Sentiment zu signifikant negativen Marktreaktionen führt. Dabei verdeutlicht die Konsistenz der Ergebnisse, dass das Sentiment in Earnings Calls verlässlich und robust die abnormalen Renditen im initialen Reaktionsfenster beeinflusst. Dies deutet darauf hin, dass das Sentiment von den Investoren systematisch verarbeitet wird und nicht nur zufällig mit den Renditen korreliert. Das Bestimmtheitsmaß fällt mit Werten zwischen 0,0630 und 0,0809 gering aus, was darauf hinweist, dass lediglich ein geringer Teil der Varianz der abnormalen Renditen durch die unabhängigen Variablen erklärt wird. Diese Tatsache ist bei Finanzmarktdaten, die in der Regel von einer Vielzahl an schwer quantifizierbaren Faktoren beeinflusst werden, nicht ungewöhnlich und konsistent zu anderen Arbeiten zu Earnings Calls.[168] Die höchste Erklärungskraft bezüglich des Sentiments weist das mit dem Henry-Lexikon ermittelte Modell auf, während das Modell des Harvard-GI Lexikons das niedrigste Bestimmtheitsmaß zeigt. Des Weiteren verzeichnet das Modell des Harvard-GI Lexikons eine höhere Standardabweichung für den Koeffizienten, als die beiden anderen Lexika, was auf eine größere Unsicherheit der Schätzung hinweist.

Der Koeffizient des Harvard-GI-Lexikons ist mit 0,281 der höchste unter den untersuchten Lexika und deutet zunächst auf einen starken positiven Effekt hin. Der hohe Wert könnte jedoch auf die geringere Spannweite der Sentiment-Werte in diesem Lexikon zurückzuführen sein, wodurch geringe Änderungen eine größere relative Auswirkung haben. Ein hoher Koeffizient indiziert eine hohe Sensitivität des Lexikons, wobei jedoch zu berücksichtigen ist, dass dies auch auf weniger differenzierte Sentiment-Werte zurückzuführen sein könnte. Diese Feststellung schmälert nicht zwangsläufig die Eignung des Lexikons, sondern demonstriert, dass der Effekt des Sentiments pro Einheit potenziell größer sein kann.

Die Sentimentvariablen erweisen sich nicht nur als signifikant, sondern zeigen zudem höhere Koeffizienten als andere bedeutsame Einflussfaktoren, wie beispielsweise die Unternehmensgröße, der Verschuldungsgrad oder das Umsatzwachstum. Andere Kontrollvariablen, wie beispielsweise der Return on Assets (ROA) sowie die Anzahl der an dem Call teilnehmenden Analysten, weisen demgegenüber

[168] Vgl. De Amicis et al. (2021); Brockman et al. (2015, p. 36); Cicon (2015, p. 75); Huang et al. (2013, p. 1006); Price et al. (2012, p. 1005).

keinen signifikanten Einfluss auf. Die Effekte des Sentiments sind insgesamt konsistenter, signifikanter und stärker als die der Kontrollvariablen. Dies demonstriert, dass die qualitativ und emotional geprägten Sentimentfaktoren in der initialen Marktreaktion einen maßgeblicheren Einfluss auf die abnormalen Renditen ausüben als rein quantitative Finanzkennzahlen.

4.2.2 Separater Einfluss von Präsentations- und Q&A-Sentiment

In einem nächsten Schritt wurde anstelle des Sentiments des gesamten Calls, das Sentiment der Präsentation $Sentiment_{Präsentation,i,q,l}$ und des Q&As $Sentiment_{Q\&A,i,q,l}$ einzeln betrachtet:

$$\begin{aligned} CAR(-1,1)_{i,q} =& \alpha + \beta_1 * Sentiment_{Präsentation,i,q,l} \\ &+ \beta_2 * Sentiment_{Q\&A,i,q,l} \\ &+ \sum Controls_{i,q} + \delta_i + \lambda_q + \varepsilon_{i,q}. \end{aligned} \tag{9}$$

Die Koeffizienten für das Sentiment der Präsentation sowie des Q&A sind durchgehend positiv und auf dem 0,1-%-Niveau signifikant (siehe Tabelle 4). Die Ergebnisse zeigen, dass sowohl das Sentiment der Präsentation als auch das Sentiment des Q&A-Abschnitts signifikante und unabhängige Einflüsse auf die abnormalen Renditen ausüben. Damit wird die Hypothese H2a gestützt. Für das LM- und Henry-Lexikon weist das Q&A leicht höhere Koeffizienten als die Präsentation auf, was auf einen stärkeren Einfluss des Frage-und-Antwort-Abschnitts auf die CAR(-1,1) hindeutet und die Hypothese 2b bestätigt. Die höheren Koeffizienten im Q&A-Abschnitt legen nahe, dass Investoren auf die spontanen Antworten des Managements oder Fragen der Analysten stärker reagieren als auf die vorbereiteten Präsentationen. Dies könnte darauf zurückzuführen sein, dass sie diese Informationen als authentischer und aufschlussreicher für die zukünftige Unternehmensentwicklung einschätzen. Im Gegensatz dazu weist das GI-Lexikon höhere Koeffizienten für den Präsentationsteil auf, was eine gegenteilige Beziehung impliziert.

Eine separate Betrachtung des Sentiments der Präsentation und des Q&As führt zu einer leichten Verbesserung des adjustierten Bestimmheitsmaßes der Modelle mit LM- und Henry-Lexikon. Diese detaillierte Betrachtung liefert folglich eine bessere Erklärung der abnormalen Renditen als das aggregierte Modell, obwohl sie eine größere Anzahl an Prädiktoren beinhaltet. Die signifikanten und differenzierten Einflüsse der einzelnen Abschnitte des Earnings Calls könnten in dem aggregierten Modell, das lediglich das Sentiment des gesamten Calls zusammenfasst, verwischt werden. Dies verdeutlicht die Relevanz einer differenzierten Betrachtung der verschiedenen Segmente von Earnings Calls. Wie bereits in der vorherigen Analyse zeigt auch diesmal das Henry-Lexikon die beste Modellgüte, während das GI-Lexikon gegenläufige Ergebnisse liefert. Des Weiteren lässt sich auch in diesem Modell anhand des Standardfehlers des Harvard-GI-Lexikons ableiten, dass die Schätzungen weniger präzise ausfallen als bei LM und Henry.

4.2.3 Differenzierter Einfluss von Manager- und Analysten-Sentiment im Q&A

Im Anschluss wurde neben der Präsentation anstelle des Sentiments des gesamten Q&As mit $Sentiment_{AnalystQ\&A,i,q,l}$ das Sentiment der Analysten und $Sentiment_{ManagerQ\&A,i,q,ö}$ das der Manager in der Frage- und Antwortsektion einzeln betrachtet:

$$\begin{aligned} CAR(-1,1)_{i,q} =& \alpha + \beta_1 * Sentiment_{Präsentation,i,q,l} \\ &+ \beta_2 * Sentiment_{ManagerQ\&A,i,q,l} \\ &+ \beta_3 * Sentiment_{AnalystQ\&A,i,q,l} \\ &+ \sum Controls_{i,q} + \delta_i + \lambda_q + \varepsilon_{i,q}. \end{aligned} \tag{10}$$

In Bezug auf die Koeffizienten des Präsentations-Sentiments lässt sich, wie aus Tabelle 5 ersichtlich, mit allen drei Lexika ein positives und hochsignifikantes Ergebnis auf dem 0,1-%-Niveau verzeichnen. Die Koeffizienten der Präsentation übersteigen die der einzelnen Elemente des Q&A-Abschnitts. Der vorherige Vergleich hat gezeigt, dass das Q&A insgesamt eine größere Bedeutung hat. Bei einer differenzierteren Betrachtung wird jedoch deutlich, dass das Sentiment in den kürzeren Segmenten der Managerantworten und Analystenfragen weniger gewichtet wird als die Präsentation.

Eine separate Betrachtung des Sentiments innerhalb des Frage-und-Antwort-Abschnitts verdeutlicht, dass das Analysten-Sentiment bei den LM- und Henry-Lexika ein hochsignifikantes Ergebnis auf dem 0,1-%-Niveau zeigt, wobei die Koeffizienten doppelt so hoch sind wie bei den Antworten des Managements, was die Hypothese H2d bestätigt. Diese Ergebnisse legen nahe, dass die höhere Bedeutung des Q&A-Abschnitts im Vergleich zur Präsentation insbesondere durch die Fragen der Analysten bedingt ist. Demgegenüber lässt sich mit dem Harvard-GI-Lexikon kein signifikanter Einfluss des Analyst-Q&A-Sentiments feststellen. Die Auswertung des Manager-Q&A-Sentiments ergibt signifikante Ergebnisse für das Harvard-GI-Lexikon (1-%-Niveau) und das LM-Lexikon (10-%-Niveau), während beim Henry-Lexikon keine Signifikanz nachgewiesen werden kann. Die vorliegenden Ergebnisse stützen die Hypothese H2c nur teilweise, da sowohl das Analysten-Sentiment als auch das Manager-Sentiment nicht durchgängig signifikante Einflüsse aufweisen.

Insgesamt deuten die Ergebnisse darauf hin, dass die Lexika verschieden auf die spezifischen und differenzierten Einflüsse der einzelnen Teile der Earnings Calls reagieren und unterschiedliche Fähigkeiten zur präzisen Erfassung der jeweiligen Segmente besitzen. Das Henry-Lexikon verzeichnet erneut die beste Modellgüte, während das Harvard GI-Lexikon ein negatives adjustiertes Bestimmtheitsmaß aufweist und somit keine höhere Erklärungskraft bietet als der Mittelwert der abhängigen Variable. Die Modellgüte für

Tabelle 5: Einfluss des Sentiments in der Präsentation sowie der Aussagen von Managern und Analysten im Q&A von Earnings Calls auf CAR(-1,1)

	LM	HE	GI	Komb.
Präsentation	*0.056****	*0.070****	*0.160****	*0.081****
	(0.0141)	*(0.0133)*	*(0.0340)*	*(0.0124)*
Manager Q&A	*0.018*	*0.018*	*0.079***	*0.062**
	(0.0106)	*(0.0127)*	*(0.0247)*	*(0.0247)*
Analyst Q&A	*0.047****	*0.044****	*0.015*	*0.049****
	(0.0065)	*(0.0082)*	*(0.0191)*	*(0.0066)*
R^2	*0.0866*	*0.0885*	*0.0688*	*0.0966*
Adj. R^2	*0.0187*	*0.0208*	*-0.0003*	*0.0295*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

die Präsentation und das Gesamtsentiment des Q&As ist für alle Lexika höher als bei einer separaten Betrachtung von Manager- und Analysten-Sentiment innerhalb des Frage-und-Antwortabschnitts. Die fehlende Signifikanz einzelner Teile des Q&As deutet darauf hin, dass die unzureichende Eignung der Lexika für bestimmte Segmente der Earnings Calls dazu führt, dass nicht signifikante Variablen einbezogen werden, die die erklärte Varianz verringern.

4.2.4 Langfristige Auswirkungen des Sentiments

Im Zeitverlauf beginnen Investoren, die veröffentlichten Informationen gründlicher zu analysieren und die langfristigen Implikationen zu bewerten, was zu einer Anpassung der Marktreaktionen führen kann. Daher wird auch der Einfluss des Sentiments auf die kumulierten abnormalen Renditen CAR(2,42) zwei bis 42 Tage nach dem Earnings Call betrachtet.

In der Driftperiode zeigt das Sentiment aller Lexika einen signifikanten negativen Einfluss auf die abnormalen Renditen, was darauf hindeutet, dass die anfängliche positive Marktreaktion nicht vollständig nachhaltig ist und eine negative Korrektur der anfänglichen Renditen eintritt (siehe Tabelle 6). Die Hypothese H1c wird somit bestätigt, wobei jedoch keine vollständige Umkehr eintritt. Diese Beobachtungen stehen im Einklang mit bestehenden Studien, die ähnliche Korrekturen belegen. So dokumentieren beispielsweise Huang et al. (2020) eine langfristige Revision der anfänglichen positiven Reaktion auf den Ton von Gewinnmitteilungen.[169] Druz et al. (2020) und Price et al. (2012) stellten hingegen fest, dass das Sentiment im Zeitverlauf an Bedeutung gewinnt und einen gleichgerichteten Drift wie die initiale Reaktion auslöst, was sie darauf zurückführen, dass qualitative Informationen schwerer zu verarbeiten sind und daher verzögert in die Marktbewertungen einfließen.[170]

Tabelle 6: Einfluss des Sentiments in der Präsentation sowie der Aussagen von Managern und Analysten im Q&A von Earnings Calls auf CAR(2,42)

	LM	HE	GI	LM	HE	GI
Gesamter Call	*−0.0617**	*−0.0958****	*−0.1744****			
	(0.0249)	*(0.0246)*	*(0.0598)*			
Präsentation				*−0.045**	*−0.081****	*−0.173****
				(0.0207)	*(0.0209)*	*(0.0523)*
Q&A				*−0.025*	*0.001*	*−0.024*
				(0.0183)	*(0.0197)*	*(0.0430)*
R^2	*0.1096*	*0.1125*	*0.1096*	*0.1147*	*0.1177*	*0.1154*
Adjusted R^2	*0.0499*	*0.0530*	*0.0500*	*0.0492*	*0.0524*	*0.0499*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

Die Modellgüte ist in der Driftperiode höher als in der initialen Reaktion, wobei dieser Anstieg primär durch Kontrollvariablen bestimmt wird und das Sentiment eine untergeordnete Rolle spielt. Während die Gewinnüberraschung in der initialen Reaktion keine signifikante Auswirkung auf die abnormalen Renditen hat, lässt sich in der anschließenden Driftperiode ein signifikanter, negativer Effekt beobachten. Zudem sind in diesem Zeitraum signifikante Einflüsse der Unternehmensgröße sowie des Return on Assets festzustellen. Die Er-

[169] Vgl. Huang et al. (2013, p. 1111).
[170] Vgl. Druz et al. (2020, pp. 59, 66); Price et al. (2012, p. 994).

gebnisse deuten darauf hin, dass nach der initialen, emotional geprägten Reaktion eine verstärkte Fokussierung auf quantitative Unternehmenskennzahlen und langfristige Bewertungen stattfindet, bei der Investoren potenzielle Risiken und Unsicherheiten intensiver berücksichtigen und ihre ursprüngliche Reaktion teilweise revidieren. Dies unterstreicht die Notwendigkeit, sowohl qualitative als auch quantitative Aspekte in die Analyse von Unternehmenskommunikationen einzubeziehen, um ein umfassendes Verständnis der Marktreaktionen zu gewährleisten.

Während das Sentiment im Rahmen des Q&A (inklusive der darin enthaltenen Aussagen von Managern und Analysten) initial eine wesentliche Rolle spielt und CAR(-1,1) beeinflusst, sind seine Auswirkungen in der späteren Driftperiode nicht signifikant. Die Präsentation hingegen zeigt einen signifikanten negativen Einfluss auf CAR(2,42) in der Driftperiode nach dem Call. Diesbezüglich lassen sich zwei Erklärungsansätze anführen: Erstens könnte dies darauf hinweisen, dass das Sentiment der Präsentation in der Driftperiode aufgrund tiefergehender Bewertungen und der Erkennung langfristiger Risiken an Bedeutung gewinnt Die Präsentation liefert wichtige Informationen zu Strategien, Wachstumsprognosen und Risiken, die anfänglich möglicherweise weniger negativ interpretiert oder erst verzögert wahrgenommen werden. Wenn Investoren die langfristigen Implikationen im Zeitverlauf gründlicher bewerten und erkennen, können negative Kursentwicklungen in der Driftperiode eintreten.

Zweitens könnte die unterschiedliche Behandlung der Informationen aus der Q&A-Sitzung und der Präsentation auf unterschiedliche Wahrnehmungen der Glaubwürdigkeit hinweisen. Q&A-Aussagen wirken oft direkter und authentischer, was sie für Investoren glaubwürdiger macht. Da sie als weniger manipulierbar angesehen werden, werden diese Informationen in der Driftperiode nicht korrigiert. Im Gegensatz dazu wird die Präsentation als strategisch und marketingorientiert betrachtet. Diese Wahrnehmung kann dazu führen, dass anfänglich positive Reaktionen auf die Präsentation in der Driftperiode relativiert werden, wenn Investoren die langfristige Tragfähigkeit und möglichen Überoptimismus in Frage stellen. Diese Korrektur zeigt sich in negativen Renditen, was darauf hinweist, dass das anfängliche Vertrauen in die Präsentation nachlässt, sobald mehr Zeit für die Analyse und Bewertung vergangen ist.

4.2.5 Vergleich der verwendeten Sentiment-Lexika

Die Unterschiede in den Ergebnissen, die sich aus der Anwendung verschiedener Lexika ergeben, verdeutlichen, dass die Wahl des Lexikons die Resultate hinsichtlich der Marktreaktionen und Interpretation der abnormalen Renditen stark beeinflusst und bestätigen die Hypothese 3a. Zusammengefasst scheinen das LM- und Henry-Lexikon besser zur Erfassung des Sentiments über den gesamten Verlauf eines Earnings Calls geeignet als das Harvard-GI Lexikon, da sie konsistentere und signifikante Ergebnisse liefern. Harvard-GI zeigt in allen Modellen die geringste Modellgüte und weist teilweise Unterschiede in den signifikanten Variablen auf, was darauf hindeutet, dass es möglicherweise weniger gut für die spezifischen Inhalte der Earnings Calls geeignet ist oder Sentiment unpräziser misst.

Die Hypothese H3b, dass das mit dem LM-Lexikon ermittelte Sentiment einen ausgeprägteren Zusammenhang zur Marktreaktion zeigt als das Sentiment, das mit den Lexika von Henry oder Harvard GI ermittelt wird, kann in dieser Form nicht bestätigt werden. Die divergierende Modellgüte der Lexika in den verschiedenen Abschnitten des Earnings Calls demonstriert, dass kein einzelnes Lexikon als universell überlegen erachtet werden kann. Die Eignung eines Lexikons ist maßgeblich von der Textart und dem spezifischen Kontext abhängig. Da die Lexika für unterschiedliche Anwendungsbereiche entwickelt wurden, variiert ihre Fähigkeit zur Sentimenterfassung in Abhängigkeit von dem jeweiligen Abschnitt des Earnings Calls. Das Harvard-GI-Lexikon wurde ursprünglich für die Sozialpsychologie erstellt und basiert auf allgemeinen psychosozialen Wortlisten.[171] In Anbetracht der nicht auf den finanziellen Kontext ausgerichteten Konzeption des Lexikons erweist sich dieses als weniger geeignet für den Präsentationsteil von Earnings Calls, in dem eine spezifische Finanzsprache zum Einsatz kommt. Das Henry-Lexikon wurde hingegen speziell für die Analyse von Gewinnmitteilungen entwickelt[172] und empfiehlt sich insbesondere für die Verwendung in Präsentationen, da diese in der Regel eine ähnliche Sprache verwenden. Auch das LM-Lexikon, das auf einer Analyse von US-Jahresberichten basiert,[173] ist in der Lage, derartige Präsentationen gut zu erfassen. Allerdings könnte die Präsentation eher einer Gewinnmitteilung als einem Jahresabschluss ähneln, was das LM-Lexikon im Vergleich zum Henry-Lexikon etwas benachteiligt.

Innerhalb des Q&As lassen sich die Aussagen der Unternehmensvertreter mit dem GI-Lexikon am adäquatesten erfassen, da es eine allgemeinere Sprache und eine höhere Affinität zu psychosozialen Inhalten aufweist. Dies ist insbesondere in einem spontanen Kontext von Vorteil. Im Vergleich zu den Formulierungen in Gewinnmitteilungen und Jahresberichten ist die von den Managern in diesem Abschnitt verwendete Sprache dynamischer und konversationsartiger, was die Anwendung der anderen beiden Lexika erschweren könnte. Die von Analysten vorbereiteten und spezifischen finanziellen Fragen lassen sich mittels finanzspezifischer Lexika besser analysieren als durch das Harvard GI-Lexikon. Infolge der umfangreicheren Wortanzahl ist davon auszugehen, dass das LM-Lexikon in diesem Abschnitt eine höhere Anpassungsfähigkeit und präzisere Messung aufweist als das Henry-Lexikon.

Die Weiterentwicklung und Anpassung der Lexika für die verschiedenen Abschnitte von Earnings Calls stellt eine vielversprechende Möglichkeit für eine Verbesserung der Sentiment-Erfassung dar. Eine Kombination mehrerer Lexika und eine Analyse jedes Abschnitts, mit dem am besten für diesen geeigneten Lexikon könnten zu einem umfassenderen und präziseren Verständnis der Stimmungslage führen.

[171] Vgl. Stone und Hunt (1963, pp. 241f).
[172] Vgl. Henry (2008, p. 384).
[173] Vgl. Loughran und McDonald (2016, p. 1204).

Zur Überprüfung dieser Annahme wird eine Regression durchgeführt, in der das Sentiment der Präsentation mit dem Henry-Lexikon, das Sentiment der Analystenfragen mit dem LM-Lexikon sowie das Sentiment der Antworten der Unternehmensleitung im Q&A des Earnings Calls mit dem Harvard GI-Lexikon ermittelt wird. Das Resultat belegt eine höhere Modellgüte im Vergleich zu allen vorherigen Modellen und signifikante Werte für Präsentation und Analystenfragen auf dem 0,1-%-Niveau sowie für Managerantworten auf dem 5-%-Niveau (siehe Tabelle 5). Während Hypothese H2c in der Betrachtung der Segmente mit einem einzelnen Lexikon nicht vollständig bestätigt werden konnte, belegt die Untersuchung mit kombinierten Lexika nun einen signifikanten und unabhängigen Einfluss des Sentiments der Manager und Analysten innerhalb des Q&A-Abschnitts auf die Marktreaktion. Zudem übertrifft dieses Modell hinsichtlich der Erklärungskraft sämtliche anderen Kombinationen der drei Lexika.

Die dargelegten Befunde verdeutlichen die Einschränkungen lexikonbasierter Sentimentmethoden, insbesondere im Hinblick auf die Erfassung von Kontext und Mehrdeutigkeit von Wörtern. Dabei weist die Quantifizierung der qualitativen Anlegerstimmung in einem Sentimentmaß inhärente Ungenauigkeiten auf.[174] Klassifikationsfehler oder die Nichterkennung spezifischer Begriffe können dabei zu einer Verzerrung der Sentimentmessung und somit zu einer Beeinträchtigung der Genauigkeit der geschätzten Regressionskoeffizienten führen.[175] Besonders in Earnings Calls, die verschiedene Abschnitte und Redner mit unterschiedlichen Themen und Stilen umfassen, erschwert dies eine präzise Sentimentanalyse und kann zu weniger verlässlichen Ergebnissen führen. Die Ungenauigkeiten in der Messung könnten die teilweise geringe Aussagekraft der Sentimentindikatoren erklären, ohne dass diese die tatsächliche Relevanz des Sentiments für Anleger reflektiert.

Angesichts dieser Einschränkungen lexikonbasierter Methoden könnte der Einsatz maschineller Lernverfahren eine vielversprechende Alternative darstellen, um die Sentimentanalyse in Earnings Calls zu präzisieren. Frankel et al. (2021) stellten fest, dass besonders in spontaneren und dynamischeren Textquellen die Vorteile maschinellen Lernens gegenüber lexikonbasierten Methoden deutlich werden.[176] Die Kenntnis spezifischer Bedeutungen und Tonfälle in unterschiedlichen Kontexten, wie verschiedenen Segmenten eines Earnings Calls, kann durch entsprechende Trainingsdaten erworben werden. In diesem Zusammenhang kann beispielsweise die thematische und stilistische Variation im Q&A-Abschnitt im Vergleich zur Präsentation bei der Klassifikation von Wörtern Berücksichtigung finden. Ein weiterer Vorteil des maschinellen Lernens besteht in der Fähigkeit, spezifische Begriffe, die in einem Lexikon möglicherweise fehlen, durch Modelltraining zu identifizieren und zu verarbeiten.

4.3 Einfluss des Sentiments in Abhängigkeit von der Informationslage

Die Reaktion der Investoren auf den Tonfall in Earnings Calls dürfte maßgeblich durch ihre jeweilige Informationslage beeinflusst werden. Dabei spielen sowohl persönliche Anlegerfaktoren als auch das Informationsumfeld des Unternehmens eine Rolle. In dieser Arbeit werden daher der Anteil institutioneller Anleger, die S&P 500-Zugehörigkeit, der ESG-Score und die Dividendenausschüttung untersucht. Da für die genannten Aspekte keine quartalsweisen Daten verfügbar sind oder die Werte über die Zeit hinweg stabil bleiben, können sie nicht direkt in die Regression einbezogen werden. Dies ist insbesondere relevant für Panelregressionen mit fixen Effekten, die Wertänderungen berücksichtigen. Das Ergebnis eines Hausman-Tests weist einen p-Wert nahe Null auf, was darauf hinweist, dass eine Random-Effects-Schätzung in diesem Fall inkonsistent wäre und das Fixed-Effects-Modell zu bevorzugen ist. Um dennoch den Einfluss des Anteils institutioneller Investoren und des ESG-Scores zu analysieren, wurden die Daten nach dem Terzil dieser Variablen unterteilt und separate Regressionen für das höchste und das niedrigste Terzil durchgeführt. Für die Zugehörigkeit zum S&P 500 und das Zahlen von Dividenden werden die Beobachtungen in Gruppen eingeteilt und beide Gruppen direkt miteinander verglichen. Für die Analyse wird jeweils das Sentiment in der Präsentation mit dem Henry-Lexikon, das der Manager Aussagen im Q&A mit dem Harvard-GI-Lexikon und das der Analystenfragen mit dem LM-Lexikon ermittelt.

4.3.1 Anteil institutioneller Anleger

Zunächst wird der Zusammenhang zwischen der Anlegerkompetenz, gemessen am Anteil institutioneller Investoren, und der Marktbewertung des linguistischen Tons in Earnings Calls untersucht. Ein hoher Anteil institutioneller Anleger lässt auf eine Beobachtung des Unternehmens durch erfahrene und fachkundige Marktteilnehmer mit umfassenden Informationsressourcen und Expertise schließen.[177] In Übereinstimmung mit den Studien von Bochkay et al. (2020) und Brockman et al. (2015) zeigen die Ergebnisse in Tabelle 7, dass Unternehmen mit einem höheren Anteil institutioneller Investoren stärkere Marktreaktionen auf das Sentiment in allen Abschnitten der Earnings Calls verzeichnen. Für institutionelle Anleger ist das Sentiment stärker signifikant, weist höhere Koeffizienten auf und zeigt eine bessere Modellgüte. Während für private Anleger vor allem die Präsentation von Relevanz ist, legen institutionelle Investoren großen Wert auf die Stimmung der Unternehmensvertreter im Q&A. Dies deutet darauf hin, dass institutionelle Investoren ihre Entscheidungen auf eine umfassendere und detailliertere Analyse der erhaltenen Informationen stützen, während private Anleger stärker auf einfach zugängliche und klar strukturierte Inhalte angewiesen sind. Die Ergebnisse legen insgesamt nahe, dass

[174] Vgl. Loughran und McDonald (2016, pp. 1189–1191).
[175] Vgl. Frankel et al. (2021, p. 5516); Loughran und McDonald (2016, p. 1226).
[176] Vgl. Frankel et al. (2021, pp. 5516–5518).
[177] Vgl. Bochkay et al. (2019, p. 50); Brockman et al. (2015, pp. 25 und 28).

Tabelle 7: Einfluss des Sentiments in Earnings Calls auf die CAR in Abhängigkeit vom Anteil institutioneller Anleger (IO)

	CAR(-1,1)		*CAR(2,42)*	
	IO gering	IO hoch	IO gering	IO hoch
Präsentation	*0.068***	*0.088****	*−0.103***	*−0.033*
	(0.0255)	*(0.0179)*	*(0.0386)*	*(0.0310)*
Manager Q&A	*0.084*	*0.108***	*0.034*	*−0.036*
	(0.0459)	*(0.0407)*	*(0.0731)*	*(0.0580)*
Analyst Q&A	*0.047****	*0.048****	*−0.025*	*0.007*
	(0.0113)	*(0.0108)*	*(0.0200)*	*(0.0151)*
R^2	*0.0931*	*0.1037*	*0.1347*	*0.1462*
Adjusted R^2	*-0.0035*	*0.0245*	*0.0425*	*0.0708*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

das Sentiment für institutionelle Investoren bedeutsamer ist und sie besser in der Lage sind, wertrelevante Informationen aus dem linguistischen Ton zu extrahieren als weniger geschulte individuelle Investoren. Letztere sind stärker auf offensichtliche, quantitative Informationen angewiesen und zeigen eine geringere Fähigkeit, subtile Nuancen im Tonfall der Teilnehmer des Earnings Calls zu interpretieren.

Diese Beobachtung zeigt sich auch in der Driftperiode: Die negative Signifikanz des Präsentationssentiments in der Driftperiode bei privaten Anlegern deutet darauf hin, dass sie diese Signale erst verzögert wahrnehmen oder anfangs weniger negativ interpretieren, was zu späteren negativen Kurskorrekturen führt. Bei Unternehmen mit einem hohen Anteil institutioneller Investoren zeigt das Präsentationssentiment in der Driftperiode hingegen keinen signifikanten Einfluss. Institutionelle Investoren bewerten die Informationen demnach bereits im initialen Reaktionsfenster genauer. Die Aussagen von Analysten und Managern während des Q&A-Abschnitts sind für beide Gruppen in der Driftperiode nicht signifikant.

4.3.2 S&P500 Zugehörigkeit

Zweitens zeigt ein Vergleich zwischen Unternehmen, die im S&P 500 gelistet sind, und solchen, die dies nicht sind, dass Investoren kleinerer Unternehmen empfindlicher auf Kommunikationsaspekte in Earnings Calls reagieren (siehe Tabelle 8). Die strengen regulatorischen Anforderungen sowie die starke öffentliche Aufmerksamkeit führen bei S&P-500-Unternehmen zu einer umfassenderen Informationsverfügbarkeit.[178] Das Sentiment in Präsentationen und Analystenfragen wirkt sich signifikant auf die abnormalen Renditen beider Gruppen aus, wobei dieser Effekt bei Unternehmen außerhalb des S&P 500 stärker ausgeprägt ist. Des Weiteren lässt sich ein signifikanter Einfluss der Aussagen von Managern während der Q&A-Sitzungen lediglich bei diesen Unternehmen feststellen. Diese Erkenntnis lässt sich durch höhere Unsicherheit und geringere Informationsverfügbarkeit erklären, weshalb Anleger stärker auf zusätzliche Informationen und Nuancen in spontanen Antworten des Managements angewiesen sind. Die Beobachtungen stehen im Einklang mit der Annahme von Milian et al. (2017), die den vergleichsweise geringen ökonomischen Einfluss des Sentiments von S&P-500-Unternehmen in ihrer Studie auf deren effiziente Bewertung zurückführen.[179]

In der Driftperiode lässt sich ein signifikant negativer Einfluss des Sentiments der Präsentation nur bei Unternehmen beobachten, die nicht dem S&P 500 angehören. Dies deutet darauf hin, dass diese Unternehmen einer höheren Anfälligkeit für nachträgliche Neubewertungen sowie einer verzögerten Erkennung langfristiger Risiken unterliegen, während bei S&P-500-Unternehmen unmittelbar eine bessere Verfügbarkeit und gründlichere Bewertung der Informationen vorliegt. Die Aussagen von Analysten und Managern während der Q&A-Sitzungen üben in der Driftperiode für Investoren beider Gruppen keinen signifikanten Einfluss auf die abnormalen Renditen aus.

4.3.3 ESG-Score

Drittens wird der Zusammenhang zwischen der Nachhaltigkeit eines Unternehmens, bewertet anhand seines ESG-Scores, und dem Einfluss des Sentiments in Earnings Calls untersucht. Der von LSEG ermittelte ESG-Score basiert auf vom Unternehmen selbst bereitgestellten Informationen zu Umwelt-, Sozial- und Unternehmensführungsaspekten und dient der Messung der relativen Nachhaltigkeit der ESG-Berichterstattung eines Unternehmens.[180] Unternehmen, die hohe ESG-Scores aufweisen, zeichnen sich durch eine gesteigerte Transparenz sowie ethischere Geschäftspraktiken aus. Dies führt zu einer Verringerung der Informationsasymmetrie zwischen Unternehmen und Investoren sowie der Agency-Probleme, da Investoren Zugang zu umfassenderen und verlässlicheren Informationen erhalten.[181] Die Ergebnisse der Regressionen in Tabelle 9 legen nahe, dass Investoren bei Unternehmen mit niedrigem ESG-Score insgesamt stärker auf das Sentiment reagieren. Eine differenzierte Betrachtung der Abschnitte in den Earnings Calls zeigt, dass diese je nach ESG-Score unterschiedlich gewichtet

[178] Vgl. H. Chen et al. (2004, p. 1928).

[179] Vgl. Milian et al. (2017, p. 73).

[180] Vgl. LSEG (2024).

[181] Vgl. Huang et al. (2024), p. 1156.

Tabelle 8: Einfluss des Sentiments in Earnings Calls auf die CAR in Abhängigkeit von der Zugehörigkeit zum S&P 500 Index

	CAR(-1,1)		*CAR(2,42)*	
	S&P500	Rest	S&P500	Rest
Präsentation	*0.075****	*0.085****	−0.029	−0.097*****
	(0.0162)	*(0.0158)*	*(0.0304)*	*(0.0243)*
Manager Q&A	*0.003*	*0.080**	−0.011	−0.013
	(0.0324)	*(0.0318)*	*(0.0629)*	*(0.0452)*
Analyst Q&A	*0.030****	*0.056****	*0.003*	*−0.009*
	(0.0085)	*(0.0083)*	*(0.0133)*	*(0.0130)*
R^2	*0.0683*	*0.1106*	*0.1228*	*0.1286*
Adjusted R^2	*-0.0244*	*0.0412*	*0.0356*	*0.0606*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

Tabelle 9: Einfluss des Sentiments in Earnings Calls auf CAR in Abhängigkeit vom ESG-Score

	CAR(-1,1)		*CAR(2,42)*	
	ESG gering	ESG hoch	ESG gering	ESG hoch
Präsentation	*0.084****	*0.111****	−0.131*****	−0.071****
	(0.0218)	*(0.0183)*	*(0.0350)*	*(0.0272)*
Manager Q&A	*0.100**	*0.033*	*0.116*	−0.038
	(0.0477)	*(0.0373)*	*(0.0763)*	*(0.0593)*
Analyst Q&A	*0.050****	*0.043****	−0.020	*0.010*
	(0.0121)	*(0.0110)*	*(0.0191)*	*(0.0159)*
R^2	*0.1112*	*0.0939*	*0.1366*	*0.1300*
Adjusted R^2	*0.0251*	*0.0069*	*0.0529*	*0.0464*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

werden. Bei Unternehmen mit niedrigem ESG-Score und geringerer Transparenz müssen Investoren alle verfügbaren Informationen, einschließlich der spontanen Aussagen im Q&A, intensiv nutzen, um Unsicherheiten zu reduzieren und fundierte Entscheidungen zu treffen. Unternehmen mit einem hohen ESG-Score genießen demgegenüber ein größeres Vertrauen, sodass Investoren sich in stärkerem Maße auf das Sentiment in den vorbereiteten Präsentationen verlassen. Die spontane Frage- und Antwortrunde besitzt in diesem Kontext keinen signifikanten Einfluss, da anzunehmen ist, dass die vorab bereitgestellten Informationen von den Investoren als verlässlich und ausreichend für fundierte Entscheidungen eingeschätzt werden. Unabhängig vom ESG-Score spielen Analystenfragen eine zentrale Rolle für Investoren aller Unternehmen.

Die Untersuchung des Einflusses des ESG-Scores auf die Investorenreaktionen während der Driftperiode ergibt, dass das Sentiment der Präsentation bei Unternehmen mit niedrigem ESG-Score einen signifikanteren und stärker negativen Einfluss auf die abnormalen Renditen hat als bei Unternehmen mit hohem ESG-Score. Die geringere Transparenz und die höheren Risiken bei diesen Unternehmen bedingen eine intensivierte, nachgelagerte Analyse durch Investoren, was zu stärkeren Anpassungen im zeitlichen Verlauf führt. Bei nachhaltigeren Unternehmen besteht für Investoren aufgrund der umfassenderen und verlässlicheren Informationen offenbar ein geringerer Bedarf, ihre anfänglichen Einschätzungen in der Driftperiode zu korrigieren. Aussagen von Analysten und Managern während der Q&A-Sitzungen erweisen sich demgegenüber für beide Gruppen in der Driftperiode als nicht signifikant.

4.3.4 Dividendenzahlungen

Abschließend stellen Dividenden einen wichtigen Indikator für die finanzielle Stabilität und die Sicherheit zukünftiger Cashflows eines Unternehmens dar.[182] Die Sentimentvariablen weisen in der initialen Reaktion bei beiden Gruppen vergleichbare Signifikanzniveaus auf. In Übereinstimmung mit Price et al. (2012) kann anhand des Vergleichs der Koeffizienten festgestellt werden, dass die Reaktion der Investoren auf die verschiedenen Aspekte des Sentiments in Earnings Calls bei Unternehmen, die Dividenden zahlen, etwas geringer ausfällt (siehe Tabelle 10). Die positive Signalwirkung der Dividenden kann demnach dazu führen, dass Investoren bei diesen Unternehmen anfänglich das Sentiment nicht umfassend berücksichtigen.

Zusammenfassend legen die Ergebnisse nahe, dass eine geringere Informationslage der Investoren eines Unternehmens zu einer stärkeren Reaktion auf das Sentiment in seinen Earnings Calls führt, was die Hypothese 1b bestätigt. Die Kompetenz der Anleger sowie das Informationsumfeld des Unternehmens spielen eine zentrale Rolle bei der Formung ihrer Erwartungen und der Verarbeitung von Informa-

[182] Vgl. Price et al. (2012, pp. 995f).

Tabelle 10: Einfluss des Sentiments in Earnings Calls auf die CAR in Abhängigkeit von dem Zahlen von Dividenden

	CAR(-1,1)		*CAR(2,42)*	
	Dividenden	Keine	Dividenden	Keine
Präsentation	*0.072****	*0.107****	−0.079∗∗∗	−0.0634
	(0.0138)	*(0.0230)*	*(0.0219)*	*(0.0392)*
Manager Q&A	*0.050*	*0.082*	−0.048	*0.028*
	(0.0281)	*(0.0437)*	*(0.0398)*	*(0.0717)*
Analyst Q&A	*0.044****	*0.057****	*0.000*	−0.014
	(0.0080)	*(0.0113)*	*(0.0107)*	*(0.0194)*
R^2	*0.0884*	*0.1128*	*0.1599*	*0.1076*
Adjusted R^2	*0.0159*	*0.0311*	*0.0931*	*0.0254*

**** p < 0.1% ** p < 1% * p < 5% . p < 10%*

tionen. Unternehmen, die einen geringen Anteil institutioneller Investoren aufweisen, nicht im S&P 500 gelistet sind, einen niedrigen ESG-Score besitzen oder keine Dividenden ausschütten, erfahren eine stärkere Reaktion der Investoren auf qualitative Informationen aus dem Tonfall in Earnings Calls. Insbesondere bei diesen Unternehmen kann die Auswertung von Texten folglich einen wertvollen Beitrag zur Finanzanalyse leisten. Für weniger geschulte individuelle Investoren stellt die Interpretation subtiler Nuancen im Tonfall der Teilnehmer des Earnings Calls eine Herausforderung dar.

Darüber hinaus lässt sich festhalten, dass auch die Relevanz der unterschiedlichen Segmente von Earnings Calls maßgeblich von der Informationslage des jeweiligen Unternehmens abhängig ist. Die Präsentation stellt im Vergleich zu den einzelnen Bestandteilen des Q&As den größten und unmittelbarsten Einfluss auf die Entscheidungsfindung dar. Für Unternehmen, die durch eine höhere Unsicherheit und geringere Transparenz gekennzeichnet sind, erlangen die spontanen Antworten des Managements im Rahmen des Q&As eine entscheidendere Bedeutung. Diese bieten den Investoren zusätzlichen Informationswert, indem sie einen Einblick in unvorhergesehene Themen und die Fähigkeit des Managements geben, spontan auf wichtige Fragen zu reagieren. Während institutionelle Investoren ihre Entscheidungen auf eine tiefere Analyse des Q&As stützen, reagieren private Anleger stärker auf die leichter zugänglichen Inhalte der Präsentation. Unabhängig von den spezifischen Unternehmensmerkmalen liefern die Analysten in dem initialen Ereignisfenster für Investoren stets wertvolle Informationen, wobei ihr Einfluss bei Unternehmen mit geringerer Informationsverfügbarkeit tendenziell stärker ausfällt. Die anfängliche Marktreaktion auf die Stimmung der Unternehmensvertreter in der Präsentation wird verstärkt bei Unternehmen mit unzureichender Informationslage korrigiert, während bei informationseffizienten Unternehmen weniger Anpassungen erforderlich sind. Insbesondere private Anleger erfassen Signale dabei oft verzögert, was zu späteren Korrekturen führt.

4.4 Robustheit

4.4.1 Alternative Ermittlung der Gewinnüberraschung

Die fehlende Signifikanz der Gewinnüberraschung in der initialen Marktreaktion ist insofern überraschend, als sie in anderen Veröffentlichungen häufig als maßgeblicher quantitativer Einflussfaktor auf abnormale Renditen identifiziert wurde.[183] Um diese Beobachtung zu untersuchen, wird zusätzlich ein weiterer etablierter Ansatz zur Messung der Gewinnüberraschung herangezogen. Im Rahmen dieses Ansatzes erfolgt ein Vergleich der tatsächlich gemeldeten Gewinne $EPS_{i,q}$ pro Aktie eines Unternehmens i im Quartal q mit den von Analysten antizipierten Gewinnen $AnalystEPS_{i,q}$:

$$AnalystSURP_{i,q} = \frac{EPS_{i,q} - AnalystEPS_{i,q}}{AnalystEPS_{i,q}} * 100. \quad (11)$$

Auch die anhand von Analystenprognosen ermittelte Gewinnüberraschung zeigt im dreitägigen Ereignisfenster um den Earnings Call keinen signifikanten Einfluss und ist im Gegensatz zur saisonalen Messung auch in der Driftperiode nicht signifikant. Des Weiteren fällt die Modellgüte in beiden Zeiträumen leicht geringer aus als bei der saisonalen Messung. Die fehlende Signifikanz der Gewinnüberraschung in der initialen Marktreaktion, gemessen durch den saisonalen und den Analystenansatz, lässt den Schluss zu, dass in dem spezifischen Kontext dieser Arbeit andere Faktoren, wie das Sentiment der Teilnehmer oder externe Einflüsse, eine wichtigere Rolle spielen. In Bezug auf das Vorzeichen und die Signifikanz der Sentimentvariablen lassen sich bei beiden Ansätzen zur Ermittlung der Gewinnüberraschung konsistente Ergebnisse beobachten, was deren Robustheit unterstreicht.

4.4.2 Alternative Ermittlung abnormaler Renditen

Wie zuvor in der Methodik dargelegt, können unterschiedliche Methoden zur Ermittlung abnormaler Renditen zu stark unterschiedlichen Ergebnissen führen.[184] Zur Prüfung der Robustheit der Ergebnisse wird daher neben dem Marktmodell auch ein faktorloses Modell analysiert, das in einer Vielzahl von Studien zum Sentiment in Earnings Calls Anwendung findet.[185] In faktorlosen Modellen

[183] Vgl. De Amicis et al. (2021); Milian et al. (2017, p. 70); Price et al. (2012, p. 998).
[184] Vgl. Gehrke (2022, p. 279).
[185] Vgl. Mayew et al. (2020, p. 316); Bochkay et al. (2019, p. 36); Milian et al. (2017, p. 70); Brockman et al. (2015, p. 26); Lee (2015, p. 235); Price et al. (2012, p. 998).

erfolgt die Ermittlung der erwarteten Rendite auf Basis festgelegter Benchmarks, wie beispielsweise anhand eines Vergleichsportfolios anderer Wertpapiere.[186] In dieser Arbeit wird die erwartete Rendite eines Unternehmens i am Tag t $E(R_{i,t})$ als mittlere Rendite $\overline{R_{P,t}}$eines Portfolios von Aktien mit dem gleichen Größen- und Buchwert-zu-Markt-Quartil wie Unternehmen i berechnet:

$$E(R_{i,t}) = \overline{R_{P,t}}. \tag{12}$$

Im Vergleich zum Marktmodell zeigt das faktorlose Modell im initialen Ereignisfenster eine leicht geringere Modellgüte, in der Driftperiode hingegen einen wesentlich geringeren Wert. Das deutet darauf hin, dass das Marktmodell insbesondere die längerfristige Entwicklung nach den Earnings Calls besser erfasst. Die nahezu identischen Sentiment-Koeffizienten in beiden Modellen in der initialen Reaktion weisen sowohl auf die Robustheit der Ergebnisse als auch auf eine konsistente Beziehung zwischen erklärenden Variablen und abnormalen Renditen hin. In der Driftperiode zeigen sich jedoch Unterschiede: Während die Sentimentvariablen des faktorlosen Modells keinen signifikanten Einfluss haben, weist das Präsentationssentiment im Marktmodell einen signifikant negativen Effekt auf. Diese Beobachtung deutet darauf hin, dass das faktorlose Modell die langfristige Beziehung zwischen dem Sentiment und den abnormalen Renditen nicht zuverlässig erfasst, möglicherweise aufgrund des fehlenden Einbezugs systematischer Marktfaktoren, die in der Driftperiode eine größere Rolle spielen könnten.

Des Weiteren ist festzustellen, dass die Gewinnüberraschung in der initialen Reaktion lediglich im faktorlosen Modell signifikant ist. Dies lässt vermuten, dass das Modell die direkte Beziehung zwischen Gewinnüberraschungen und abnormalen Renditen direkter erfasst, da es weniger durch den allgemeinen Markttrend beeinflusst wird. Der Marktindex könnte die Signifikanz der Gewinnüberraschung abschwächen, indem er die allgemeine Marktstimmung einbezieht, die auch auf andere Informationen reagiert.

Insgesamt bestätigen die Ergebnisse die Robustheit der gefundenen Effekte und zeigen, dass unterschiedliche Ansätze zu der Ermittlung abnormaler Renditen zu variierenden, aber in den Kernbefunden konsistenten Ergebnissen führen.

5 Fazit

Die theoretische Betrachtung dieser Arbeit hat gezeigt, dass qualitative Informationen, insbesondere im Rahmen von Earnings Calls, in der Finanzanalyse eine zunehmend wichtige Rolle spielen. Diese Form der Unternehmenskommunikation stellt eine dynamische Plattform zur Verfügung, auf der über die reine Vermittlung von Zahlen hinausgehende, wertvolle emotionale Hinweise vermittelt werden. Die textbasierte Sentimentanalyse stellt dabei ein zentrales Instrument dar, um diese sprachlichen Nuancen systematisch zu erfassen und zu interpretieren. Sie ermöglicht ein besseres Verständnis der Auswirkungen der Stimmung von Unternehmensvertretern und Analysten auf die Marktreaktionen und leistet damit einen wesentlichen Beitrag zur Vertiefung des Wissens über die Marktpsychologie und die Informationsverarbeitung von Investoren.

Basierend auf den theoretischen Erkenntnissen zur Rolle des Sentiments in der Finanzanalyse bestätigt die empirische Untersuchung, dass emotionale und sprachliche Nuancen in Unternehmensoffenlegungen entscheidende Signale für Marktteilnehmer darstellen. Investoren betrachten das Sentiment in Earnings Calls als wichtige Informationsquelle und integrieren es fortlaufend in ihre Entscheidungsprozesse. Nach einer anfänglich emotional geprägten Reaktion bewerten die Investoren zunehmend die langfristigen Implikationen der Informationen, was in der Driftperiode oft zu einer negativen Korrektur der anfänglich positiven Marktreaktionen führt. Die Ergebnisse legen dabei insgesamt nahe, dass eine differenzierte Herangehensweise an die Sentimentanalyse notwendig ist, um die komplexen Zusammenhänge zwischen den verschiedenen Segmenten eines Earnings Calls und den damit verbundenen Marktreaktionen umfassend zu erfassen. Das Sentiment in dem Q&A-Abschnitt ist für die initialen Marktreaktionen bedeutender als die Präsentation. Innerhalb des Q&As zeigen die Fragen der Analysten zudem einen größeren Einfluss auf die Marktreaktionen als die Antworten des Managements. Während die Präsentation im zeitlichen Verlauf an Bedeutung gewinnt und zu einer Revision der anfänglich positiven Reaktion führt, bleiben die authentischeren Informationen aus dem Q&A-Abschnitt stabiler und weniger anfällig für Korrekturen.

Ferner wurde deutlich, dass die Reaktionen der Investoren auf das Sentiment in Earnings Calls in hohem Maße von ihrer Informationslage abhängig sind. Unternehmen, die einen geringen Anteil institutioneller Investoren aufweisen, nicht dem S&P 500 angehören, niedrige ESG-Scores besitzen oder keine Dividenden ausschütten, erfahren intensivere Marktreaktionen auf den Tonfall in den Earnings Calls. Im Verlauf der Driftperiode wird die anfänglich positive Marktreaktion in diesen Fällen zudem häufiger revidiert. Besonders bei Unternehmen mit höherer Unsicherheit und begrenztem Informationszugang bietet eine Analyse des Sentiments somit wertvolle Einblicke. Erfahrene institutionelle Anleger sind in der Lage, den linguistischen Ton in Earnings Calls präziser zu interpretieren und daraus gezielt Informationen für ihre Investitionsentscheidungen zu extrahieren. Auch die Bedeutung der einzelnen Segmente von Earnings Calls variiert mit der Informationslage: Die spontanen Antworten des Managements im Q&A gewinnen bei höheren Informationsasymmetrien gegenüber der Präsentation an Gewicht. Zudem scheinen professionelle Anleger besonders gut in der Lage zu sein, wertvolle Informationen aus dem Manager-Sentiment in dem Q&A zu extrahieren, während weniger erfahrene Investoren stärker auf die Präsentation fokussiert sind. Unabhängig von unternehmensbezogenen Merkmalen oder der Erfahrung der Investoren ist die Stimmung der Analysten stets von entscheidender Bedeutung.

[186] Vgl. Gerpott (2009, p. 216).

Darüber hinaus wurde ein besonderer Fokus auf die methodischen Herausforderungen einer lexikonbasierten Sentimentanalyse gelegt, um die Potenziale und Grenzen dieser Methode im Bereich der Finanzanalyse besser zu verstehen. Die lexikonbasierte Sentimentanalyse hat sich als nützlich erwiesen, allerdings zeigten sich Einschränkungen bei der Erfassung von Kontext und Mehrdeutigkeit, was die Genauigkeit der Sentiment-Messung beeinträchtigt. Insbesondere die Wahl des Lexikons ist zentral für präzise Ergebnisse, wobei keines der in dieser Arbeit verwendeten Lexika universell überlegen war. Stattdessen hängt die Eignung stark vom spezifischen Abschnitt des Earnings Calls ab. Die Arbeit legt nahe, dass eine kontextabhängige Auswahl und Kombination von Lexika eine Verbesserung der Ergebnisse bieten kann. Dabei verdeutlichen die aufgezeigten Herausforderungen, dass der Einsatz maschineller Lernverfahren einen vielversprechenden Ansatz zur Verbesserung der Präzision der Sentimentanalyse in Earnings Calls darstellen könnte.

Insgesamt liefert die Arbeit somit wichtige Erkenntnisse darüber, wie qualitative Informationen aus Earnings Calls systematisch erfasst und für die Finanzanalyse genutzt werden können. Allerdings bedarf die komplexe Dynamik zwischen den verschiedenen Informationsquellen und Analysemethoden weiterer Untersuchung. Auf Grundlage der Ergebnisse lassen sich vier zentrale Ansatzpunkte für zukünftige Forschungen identifizieren: Erstens sollte die Weiterentwicklung von Sentiment-Analysemethoden im Vordergrund stehen, wobei die Integration maschinellen Lernens vielversprechende Möglichkeiten bietet, um Stimmungen in den unterschiedlichen Segmenten der Earnings Calls präziser zu erfassen. Zweitens reduziert die Analyse von textuellen Transkripten im Vergleich zu Videoübertragungen zwar die Komplexität, kann jedoch auch mit einem möglichen Informationsverlust verbunden sein. Zukünftige Untersuchungen sollten daher auch erforschen, wie nonverbale Faktoren wie Mimik, Gestik und Körpersprache in die Sentimentanalyse einbezogen werden können.

Drittens sollte sich zukünftige Forschung intensiver mit dem Einfluss der Informationslage der Investoren auf die Interpretation und Reaktion der Investoren auf das Sentiment in Earnings Calls befassen. In diesem Zusammenhang scheint es empfehlenswert, sowohl das Informationsumfeld des Unternehmens, beispielsweise durch eine Segmentierung nach Branchen oder Unternehmen in stark regulierten Märkten, als auch persönliche Anlegerfaktoren wie die geografische Herkunft der Investoren zu untersuchen. Viertens erfordert die Sentimentanalyse eine differenzierte Herangehensweise, um die Zusammenhänge zwischen den Abschnitten und Aussagen verschiedener Redner in Earnings Calls sowie die daraus resultierenden Marktreaktionen besser zu verstehen. Insbesondere die Aussagen von CEOs und CFOs sollten inhaltlich und methodisch, etwa durch den Einsatz spezifischer Lexika, differenziert analysiert werden. Dies könnte auch den Einfluss von Agency-Problemen und persönlichen Merkmalen der Sprecher tiefer beleuchten und somit aufzeigen, wie unterschiedliche Perspektiven und potenzielle Interessenkonflikte die Wahrnehmung der Investoren beeinflussen.

In einer zunehmend datengetriebenen Finanzwelt bleibt es entscheidend, die komplexen Wechselwirkungen zwischen qualitativen Informationen und Marktreaktionen weiter zu erforschen. Die Fähigkeit, Stimmungen und psychologische Signale in Unternehmensveröffentlichungen wie Earnings Calls korrekt zu deuten, könnte für Investoren einen entscheidenden Vorteil bei der Unternehmensbewertung und der Vorhersage von Kursbewegungen darstellen. Damit wird die präzise Auswertung von Texten zu einem unverzichtbaren Bestandteil der modernen Finanzanalyse.

Literatur

Agrawal, S., Azar, P. D., Lo, A. W., & Rai, T. S. (2018). Momentum, Mean-Reversion, and Social Media: Evidence from StockTwits and Twitter. *The Journal of Portfolio Management*, *44*(7), 85–95.

Ajinkya, B. B., & Gift, M. J. (1984). Corporate Managers' Earnings Forecasts and Symmetrical Adjustments of Market Expectations. *Journal of Accounting Research*, *22*(2), 425–444.

Allee, K. D., & Deangelis, M. D. (2015). The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion. *Journal of Accounting Research*, *53*(2), 241–274.

Altman, A. M. (2019). Petition for Rulemaking to Require Filers to Submit Earnings Conference Call Transcripts on Form 8-K, Pursuant to Regulation FD Compliance. Verfügbar 17. Mai 2024 unter https://www.sec.gov/files/rules/petitions/2019/petn4-742.pdf

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, *59*(3), 1259–1294.

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance*, *72*, 132–147.

Baddeley, M. (2018). *Behavioural Economics and Finance* (Second). Routledge.

Bai, C., Duan, Y., Fan, X., & Tang, S. (2023). Financial market sentiment and stock return during the COVID-19 pandemic. *Finance Research Letters*, *54*, 103738.

Baker, H. K., & Nofsinger, J. R. (2010). *Behavioral Finance: Investors, Corporations, and Markets*. John Wiley & Sons.

Baker, M., & Wurgler, J. (2007). Investor Sentiment in the Stock Market. *The Journal of Economic Perspectives*, *21*(2), 129–151.

Ball, R., & Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, *6*(2), 159–178.

Bassyouny, H., Abdelfattah, T., & Tao, L. (2022). Narrative disclosure tone: A review and areas for future research. *Journal of International Accounting, Auditing and Taxation*, *49*, 100511.

Beyer, A., Cohen, D. A., Lys, T. Z., & Walther, B. R. (2010). The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics*, *50*(2-3), 296–343.

Blau, B. M., DeLisle, J. R., & Price, S. M. (2015). Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. *Journal of Corporate Finance*, *31*, 203–219.

Bochkay, K., Hales, J., & Chava, S. (2019). Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls. *The Accounting Review*, *95*(2), 31–60.

boerse.de. (n. d.). Was ist das Sentiment? Verfügbar 17. Mai 2024 unter https://www.boerse.de/grundlagen/charttechnik/Was-ist-das-Sentiment

Borochin, P. A., Cicon, J. E., DeLisle, R. J., & Price, S. M. (2018). The effects of conference call tones on market perceptions of value uncertainty. *Journal of Financial Markets*, *40*, 75–91.

Bowen, R. M., Davis, A. K., & Matsumoto, D. A. (2002). Do conference calls affect analysts' forecasts? *The Accounting Review*, *77*(2), 285–316.

Brockman, P., Li, X., & Price, S. M. (2015). Differences in Conference Call Tones: Managers vs. Analysts. *Financial Analysts Journal*, *71*(4), 24–42.

CFI Education Inc. (2024). Earnings call. Verfügbar 10. Mai 2024 unter https://corporatefinanceinstitute.com/resources/valuation/earnings-call/

Chen, H., Noronha, G., & Singal, V. (2004). The Price Response to S&P 500 Index Additions and Deletions: Evidence of Asymmetry and a New Explanation. *The Journal of Finance*, *59*(4), 1901–1929.

Chen, X., & Tian, Y. (2018). Learning to Perform Local Rewriting for Combinatorial Optimization. https://doi.org/10.48550/arXiv.1810.00337

Chen, Y., Han, D., & Zhou, X. (2023). Mining the emotional information in the audio of earnings conference calls: A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. *International Review of Financial Analysis*, *88*, 102704.

Cicon, J. (2015). Say it again Sam: the information content of corporate conference calls. *Review of Quantitative Finance and Accounting*, *48*(1), 57–81.

Collins, D. W., Nguyen, N. Q., & Nguyen, T. T. (2024). Manager sentiment and conditional conservatism. *Journal of Business Finance & Accounting*.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, *53*(9), 1375–1388.

D'Augusta, C., & DeAngelis, M. D. (2019). Tone Concavity around Expected Earnings. *The Accounting Review*, *95*(1), 133–164.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2014). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, *20*(2), 639–673.

Davis, A. K., & Tama-Sweet, I. (2012). Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A. *Contemporary Accounting Research*, *29*(3), 804–837.

De Amicis, C., Falconieri, S., & Tastan, M. (2021). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance*, *71*, 101809.

DeBoskey, D. G., Luo, Y., & Zhou, L. (2018). CEO power, board oversight, and earnings announcement tone. *Review of Quantitative Finance and Accounting*, *52*(2), 657–680.

Druz, M., Petzev, I., Wagner, A. F., & Zeckhauser, R. J. (2020). When Managers Change Their Tone, Analysts and Investors Change Their Tune. *Financial Analysts Journal*, *76*(2), 47–69.

Elliott, W. B., Loftus, S., & Winn, A. (2023). To read or to listen? Does disclosure delivery mode impact investors' reactions to managers' tone language? *Contemporary Accounting Research*, *41*(1), 7–38.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2009). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953.

Frankel, R., Jennings, J., & Lee, J. (2021). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science*, *68*(7), 5514–5532.

Frankel, R., Johnson, M., & Skinner, D. J. (1999). An Empirical Examination of Conference Calls as a Voluntary Disclosure Medium. *Journal of Accounting Research*, *37*(1), 133–150.

Fu, X., Wu, X., & Zhang, Z. (2019). The Information Role of Earnings Conference Call Tone: Evidence from Stock Price Crash Risk. *Journal of Business Ethics*, *173*(3), 643–660.

Gehrke, M. (2022). *Angewandte empirische Methoden in Finance & Accounting* (2. Aufl.). Springer Gabler.

Gerpott, T. J. (2009). Ereignisstudie. In C. Baumgarth, M. Eisend & H. Evanschitzky (Hrsg.), *Empirische Mastertechniken. Eine anwendungsorientierte Einführung für die Marketing und Managementforschung* (S. 203–235). Gabler.

Giancarlo, C. H. (2024). *Pure Storage, Inc. (PSTG) Q4 2024 Earnings Call Transcript*. Verfügbar 17. Mai 2024 unter https://seekingalpha.com/article/4674588-pure-storage-inc-pstg-q4-2024-earnings-call-transcript

Gregory, R. P. (2021). What determines Manager and Investor Sentiment? *Journal of Behavioural and Experimental Finance*, *30*, 100499.

Gupta, R. (2019). Manager sentiment and stock market volatility. *Journal of Management Information and Decision Sciences*, *22*(1), 11–21.

Hales, J., Kuang, X. (, & Venkataraman, S. (2011). Who Believes the Hype? An Experimental Examination of How Language Affects Investor Judgments. *Journal of Accounting Research*, *49*(1), 223–255.

Hennig, J. C., Firk, S., & Wolff, M. (2023). Credibility Signals from Soft Information: Evidence from Investor Reactions to Tone in Earnings Conference Calls. *European Accounting Review*.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363–407.

Henry, E., & Leone, A. J. (2016). Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *The Accounting Review*, *91*(1), 153–178.

Hollander, S., Pronk, M., & Roelofsen, E. (2010). Does Silence Speak? An Empirical Analysis of Disclosure Choices During Conference Calls. *Journal of Accounting Research*, *48*(3), 531–563.

Holler, J. (2016). *Einführung in die Event Study Methodik*. Shaker Verlag.

Huang, X., Teoh, S. H., & Zhang, Y. (2013). Tone management. *The Accounting Review*, *89*(3), 1083–1113.

Irani, A. J. (2004). The Effect of Regulation Fair Disclosure on the Relevance of Conference Calls to Financial Analysts. *Review of Quantitative Finance and Accounting*, *22*(1), 15–28.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, *132*(1), 126–149.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Klevak, J., Livnat, J., & Suslava, K. (2024). Gender difference in CFO communication: Evidence from earnings calls. *Journal of Accounting and Public Policy*, *44*, 107193.

Lee, J. (2015). Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts during Earnings Conference Calls. *The Accounting Review*, *91*(1), 229–250.

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2. Aufl.). Cambridge University Press.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

LSEG. (2024). LSEG ESG Scores. Verfügbar 17. August 2024 unter https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores

Luo, R., & Ye, Y. (2023). Pressure from words: The tone of investors in Chinese earnings communication conferences and managerial myopia. *Accounting & Finance*, *64*(1), 833–868.

Luo, Y., & Zhou, L. (2017). Managerial ability, tone of earnings announcements, and market reaction. *Asian Review of Accounting*, *25*(4), 454–471.

Malkiel, B. G., & Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417.

Mather, P., Ranasinghe, D., & Unda, L. A. (2021). Are gender diverse boards more cautious? The impact of board gender diversity on sentiment in earnings press releases. *Journal of Contemporary Accounting & Economics*, *17*(3), 100278.

Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions. *The Accounting Review*, *86*(4), 1383–1414.

Mayew, W. J., Sethuraman, M., & Venkatachalam, M. (2020). Individual Analysts' Stock Recommendations, Earnings Forecasts, and the Informativeness of Conference Call Question and Answer Sessions. *The Accounting Review*, *95*(6), 311–337.

Milian, J. A., Smith, A. L., & Alfonso, E. (2017). An Investigation of Analysts' Praise of Management During Earnings Conference Calls. *Journal of Behavioral Finance*, *18*(1), 65–77.

Müller, S. (2023). Expected Return Models: Choosing the Right Approach. Verfügbar 24. Mai 2024 unter https://eventstudy.de/models/expected_return.html

Ning, Y., Yang, S., & Xiao, Z. (2023). Manager sentiment, stock return, and the evolving information environment in post-IPO firms. *Journal of Economics and Finance*.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577.

Pouryousof, A., Nassirzadeh, F., & Askarany, D. (2023). Inconsistency in managers' disclosure tone: the signalling perspective. *Risks*, *11*(12), 205.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992–1011.

Rahman, S. (2023). Narrative tone and earnings persistence. *Journal of International Accounting, Auditing and Taxation*, *52*, 100562.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25–40.

Rodriguez, A. (2019). Netflix plunges more than 10% after a huge miss on subscriber growth during Q2. Verfügbar 18. Juni 2024 unter https://www.businessinsider.com/live-netflix-q2-2019-earnings-subscriber-growth-analysis-2019-7

Rogers, J. L., Van Buskirk, A., & Zechman, S. L. (2011). Disclosure Tone and Shareholder Litigation. *The Accounting Review*, *86*(6), 2155–2183.

S&P Global. (2024). *S&P Composite 1500®*. Verfügbar 17. Mai 2024 unter https://www.spglobal.com/spdji/en/indices/equity/sp-composite-1500/

Stone, P. J., & Hunt, E. B. (1963). A Computer Approach to Content Analysis, Studies Using the General Inquirer System. *Managing Requirements Knowledge, International Workshop On*, *1*, 241–256.

Tama-Sweet, I. (2014). Changes in earnings announcement tone and insider sales. *Advances in Accounting*, *30*(2), 276–282.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Todd, A., Bowden, J., & Moshfeghi, Y. (2024). Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *31*(1), 1–13.

Tuffery, S. (2023). *Deep Learning: From Big Data to Artificial Intelligence with R*. John Wiley & Sons.

Yamamoto, R., Kawadai, N., Kurita, M., & Baba, S. (2022). Managements' tone strategies by earnings call transcripts in the global markets. *Journal of Asset Management*, *23*(3), 246–255.

Yan, S., Hwang, I., Jung, H.-R., & Jung, T. (2023). Abnormal Disclosure Tone in MD&A and Analysts' Earnings Forecast: Evidence from China. *Emerging Markets Finance and Trade*, 1–17.

Yang, B., Liu, Y., & Su, Y. (2022). Earnings communication conferences and post-earnings-announcement drift: Evidence from China. *Accounting & Finance*, *63*(2), 2145–2185.

Zhou, G. (2018). Measuring investor sentiment. *Annual Review of Financial Economics*, *10*, 239–259.

Abkürzungsverzeichnis

BM Buchwert-zu-Markt

CAR(-1,1) Kumulierte abnormale Renditen im dreißigsten Fenster um den Earnings Call

CAR(-2,42) Kumulierte abnormale Renditen in der Driftperiode 2 bis 42. Tage nach dem Earnings Call

CEO Chief Executive Officer

CFO Chief Financial Officer

EMH Effizienzmarkthypothese

ESG Environmental, Social, and Governance (englisch für: Umwelt, Soziales und Unternehmensführung)

GI Harvard-GI-Lexikon

HE Henry-Lexikon

IO Anteil institutioneller Anleger

LM Loughran-and-McDonald-Lexikon

PEAD Post-Earnings Announcement Drift (englisch für: Drift nach der Gewinnankündigung)

Q&A Frage- und Antwortabschnitt eines Earnings Calls

Reg FD Regulation Fair Disclosure

ROA Return on Assets

Symbolverzeichnis

$\hat{\alpha}_i$ Geschätzte Konstante für die erwartete Rendite der Aktie i

$AnalystEPS_{i,q}$ Von Analysten erwarteter Gewinn pro Aktie des Unternehmens i im Quartal q

$ANALYSTS$ Anzahl am Earnings Call teilnehmender Analysten

$AnalystSURP_{i,q}$ Analysten-basierte Gewinnüberraschung des Unternehmens i im Quartal q

$AR_{i,t}$ Abnormale Rendite der Aktie i am Tag t

$\hat{\beta}_i$ Geschätzte Sensitivität der Aktie i gegenüber der Marktrendite $R_{m,t}$

BM Buchwert-zu-Markt-Verhältnis

$CAR(t_1, t_2)$ Kumulierte abnormale Renditen von t1 bis t2 nach dem Earnings Call des Unternehmens i im Quartal q

$\sum Controls_{i,q}$ Vektor von Kontrollvariabeln auf Unternehmensebene im Quartal q

δ_i Feste Effekte für das Unternehmen i

$EPS_{i,q}$ Gewinn je Aktie des Unternehmens i im Quartal q

$\varepsilon_{i,t}$ Störterm der Regression für Unternehmen i im Quartal q

$EPS_{i,q-4}$ Gewinn je Aktie des Unternehmens i im gleichen Quartal des Vorjahres

$E(R_{i,t})$ Erwartete Rendite der Aktie i am Tag t

$FIRMAGE$ Jahre seit dem Börsengang

$GROWTH$ Prozentuales Umsatzwachstum

i Unternehmen

l Lexikon (LM; Henry; Harvard-GI)

λ_q Feste Effekte für das Quartal q

$LEVERAGE$ Verschuldungsgrad

$LOSS$ Dummy-Variable für Verluste

$N_{negativ}$ Anzahl negativ klassifizierter Wörter

$N_{positiv}$ Anzahl positiv klassifizierter Wörter

q Quartal

$R_{i,t}$ Tatsächliche Rendite der Aktie i am Tag t

$R_{m,t}$ Rendite des Marktportfolios am Tag t

ROA Return on Assets

$\bar{R}_{p,t}$ Mittlere Rendite eines Portfolios von Aktien mit dem gleichen Größen- und Buchwert-zu-Markt-Quartil wie Unternehmen i

$Schlusskurs_{i,t}$ Schlusskurs der Aktie i am Tag t

$Schlusskurs_{i,t-1}$ Schlusskurs der Aktie am Vortag

$Schlusskurs_{i,q-4}$ Schlusskurs der Aktie des Unternehmens i am Ende des Quartals des Vorjahres

$Sentiment_{i,q}$ Sentiment

$Sentiment_{Analyst(Q\&A,i,q)}$ Sentiment der Analysten in dem Q&A

$Sentiment_{Gesamt(i,q)}$ Sentiment in dem gesamten Earnings Call

$Sentiment_{Manager(Q\&A,i,q)}$ Sentiment der Manager in dem Q&A

$Sentiment_{Prsentation(i,q)}$ Sentiment in der Präsentation

$Sentiment_{Q\&A(i,q)}$ Sentiment in dem Q&A

$SIZE$ Logarithmierte Marktkapitalisierung in Millionen US-Dollar

$SURP_{i,q}$ Gewinnüberraschung des Unternehmens i im Quartal q

t Zeit

Junior Management Science

www.jums.academy
ISSN: 2942-1861

The Impact of Biodiversity Risk on Banks' Credit Default Swap Spread Changes

Moritz Praetz

Technical University of Munich

Abstract

This paper explores the impact of biodiversity risk on banks' credit risk using a news- based biodiversity index and CDS of 39 global banks between 2015 and 2023. Using a linear OLS regression, this paper finds evidence for a significant positive relationship between biodiversity news and CDS spread changes, where negative news leads to increasing CDS prices. Furthermore, cross-sectional analyses are conducted to test for heterogeneity. Using the Kunming Declaration in 2021 as an external shock, this paper finds evidence that the relationship persists for the period after the Kunming Declaration, suggesting no significant effect of biodiversity risk before. Further tests reveal no significant impact of a country's state of biodiversity. In contrast, since the Kunming Declaration, the relationship is stronger for banks which openly disclose biodiversity risks. Banks located in the USA, the only UN nation which is not a member of the CBD, experience a weaker effect of biodiversity news on CDS spread changes. These results show that banks are subject to biodiversity-related credit risks, where expectations of new policies and regulation following the Kunming Declaration significantly affect banks' CDS spreads.

Keywords: biodiversity; banking; credit default swaps; credit risk

1 Introduction

> "We are all asset managers."
>
> *- Partha Dasgupta (2021, p. 35)*

One of the most important assets accessible to humanity is nature, with all its living and non-living components (IPBES, 2019). In addition to being the foundation of life, nature also plays a key role in the way humans interact with each other and its surroundings. Humanity has learned to use nature and its assets for its benefit, forming economies like agriculture, medicine, or construction (WEF, 2020). Such industries strongly rely on the variety of ecosystems, or in other words, life and its diversity – biodiversity. A study by the World Economic Forum (WEF) has underlined this reliance, finding that more than 50% of the world's value creation is dependent on nature-related services (WEF, 2020). But global ecosystems and their diversity are fading. Among plants and land-based animals, it is estimated that about 20% of species could be extinct in the next decades alone (Dasgupta, 2021). Consequently, this area of risk has experienced increasing attention in recent years.

In 2019, IPBES published the global assessment report on biodiversity and ecosystem services, which aims to provide policymakers and private sectors with credible information about the current state of biodiversity (IPBES, 2019). Importantly, they discuss the interaction between humanity and its environment to highlight its effects on nature's deterioration. They state that due to human action, 75% of the global land surface has been altered, 85% of wetland areas have been lost, and global forest area is heavily declining. Furthermore, the rate of species extinction is 10 to 100 times larger than on average over the past 10 million years. Key drivers of these developments, like land- and sea-use change or climate change, are caused by human interactions with the environment.

The WEF (2020) report on nature lays out the dependence of global economies on biodiversity and examines the outcomes of its loss. The risks associated with biodiversity degradation can directly or indirectly appear through different channels, all with the potential to destabilize businesses and markets. However, assessing which industries are most likely affected is not always evident. The report states that in many industries, seemingly only a small portion of value creation is highly dependent on ecosystem services. However, there exist hidden dependencies, where large parts of the gross value added are at least moderately reliant on nature. For secondary and tertiary industries, nature-related risks can be significant through supply chains, even though they don't seem to be highly dependent at first glance. The report adds that the gravity of potential losses can be shown, despite these difficulties. The food and beverages, as well as construction and agriculture industries, are highly reliant on nature services. Together they generate about 8 trillion dollars of gross value added, which could significantly decrease, as these industries suffer from nature loss.

I would like to thank my supervisor, Dr. Aida Cehajic, for her incredible support and guidance during the process of writing this paper.

DOI: https://doi.org/10.5282/jums/v11i1pp74-106

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

Given the potential severity of economic consequences, there have been approaches to motivate global solutions. In 1993, the Convention on Biological Diversity (CBD) was launched with the goal to drive biodiversity conservation and ensure fair and sustainable use of resources (CBD Secretariat, 2011). With its nearly 200 member states, coordinated steps are decided and biodiversity targets are formulated. Recently, the particularly important Kunming Declaration in 2021 has received global attention and is viewed as the biodiversity equivalent to the climate-focused Paris Agreement (Garel et al., 2024). The CBD Secretariat (2021a) has even declared it to be the "most important United Nations Biodiversity event of this decade" (p. 6). In the context of the declaration, the CBD Secretariat (2021a) has also stressed the significant role of the financial sector in working towards positive biodiversity outcomes and discussed potential risks. While financial firms can provide capital to drive the conservation of biodiversity, they also need to acknowledge how biodiversity risks can negatively affect their business. As stated by Dasgupta (2021), the "loss of biodiversity [...] results in greater volatility and uncertainty around the goods and services ecosystems provide" (p. 417). These increased risks especially influence banks' operations and credit business, which may be affected by either transition risks through new regulation, or physical risks, like the loss of species or ecosystems their clients depend on (CBD Secretariat, 2021a).

While the threat of biodiversity loss is well established, the effects on the financial sector are just starting to be studied (Calice et al., 2023). Karolyi and Tobin-de la Puente (2023) acknowledge the lack of biodiversity-related finance literature and call for research. Within the finance industry, banks assume a special role due to their business models, asset-liability structures, and regulatory requirements. Consequently, research often excludes them from empirical credit risk investigations (Hasan et al., 2016). However, banks' risks are significantly exposed to biodiversity concerns through several channels, according to Hudson (2024). He explains that biodiversity loss could disrupt supply chains and harm the operating business of banks' clients, causing potential default on loan obligations. Furthermore, the value of security portfolios may decline as economies struggle with the consequences of biodiversity loss, which potentially leads to downgrades of a bank's creditworthiness. Regulatory issues, resulting in legal or reputational repercussions, or the loss of potential investment opportunities due to biodiversity concerns, add to the risks banks face (Hudson, 2024). This thesis aims to answer the call for research by Karolyi and Tobin-de la Puente and investigates whether biodiversity risks significantly influence banks' credit risk.

To examine this relation, a sample of 39 global banks is constructed for the time between 2015 and 2023. As a measure of biodiversity risk, this thesis uses a novel approach published by Giglio et al. (2023). They construct a news-based measure by analyzing New York Times articles for specific keywords related to biodiversity. Relevant reports are classified regarding their sentiment to aggregate positive and negative news. This creates a high-frequency biodiversity index, which provides daily data observations. As a measure for credit risk, credit default swaps (CDS) are used. These derivatives act as insurance in case of credit events and therefore provide a measure of credit default risk of the underlying entity (Zhu, 2006). In addition, control variables are used, which have shown to be significant determinants of CDS prices in credit derivative literature. One set of controls are balance sheet variables, including leverage ratio, asset quality, bank size, funding stability, cost efficiency, and sensitivity to market risk. Additional non-balance sheet variables include a bank's stock volatility as well as returns and volatility of stock indices. Lastly, sovereign bond yields and credit ratings are included in the analysis.

The analysis employs a linear panel ordinary least squares (OLS) estimation with fixed effects and clustered standard errors. The baseline regression provides evidence for a significant positive relationship between biodiversity news and CDS spread changes. The economic magnitude of the effect is in the range of a 1.532% to 1.563% increase in CDS price changes for every standard deviation increase of the biodiversity index. Beside biodiversity, the equity variables are all significantly affecting CDS price changes as well. Balance sheet variables do not have a significant impact, which persists when introducing a time lag of one period. To test for heterogeneity of this relationship, further analyses introduce cross-sectional interaction terms. Motivated by its significance and public attention, this thesis studies the effect of the Kunming Declaration as an external shock. For the full sample, the analysis only shows evidence for a significant positive effect in the post-Kunming period and not before. These results are in line with the findings of Garel et al. (2024), who find similar biodiversity risk effects in stock markets only after the declaration. Introducing a country's state of biodiversity does not show any diverging effects compared to the baseline regression. For the banks in the sample, the country's biodiversity degradation therefore does not seem to influence the magnitude of the relation. Kölbel et al. (2022) and Carbone et al. (2021) show that the disclosure and acknowledgement of climate-related risk has been shown to have impacts on firm's credit risk. Similarly, this thesis includes the participation of banks in responsible banking initiatives as so-called green banks to test for heterogeneity. For the base interaction, there is no evidence for a significant influence across subgroups. Including a three-way interaction to additionally test differences before and after the Kunming Declaration yields the following results. In the pre-Kunming period, green banks experienced lower influences of biodiversity news on CDS prices. In contrast, the impact is stronger for green banks in the post-Kunming period. These results are in line with the uncertainty reduction and risk perception effect acknowledged by Kölbel et al. (2022), before and after the declaration, respectively. Furthermore, the impact on US banks is studied separately, as the USA is the only United Nations (UN) member, which is not part of the CBD. Over the full sample, there is no evidence for diverging effects between US and non-US banks. How-

ever, before the Kunming Declaration, US-banks experienced increasing effects, which switch to decreasing impacts on the biodiversity-CDS relation in the post-Kunming period. The expectation of less stringent policies and regulations explains the lower impact on CDS spread changes in the post-Kunming period. The increasing effect before the declaration is not yet observed in the literature but might be caused by excessive financial distress due to natural disasters in the USA in the late 2010s. The cross-sectional differences before and after the Kunming Declaration are subject to potential multicollinearity concerns, which might affect regression outcomes. Therefore, they need to be treated with caution and are subject to further confirmation. More information will be provided in Chapter 6.2. To provide robustness, different tests are undertaken which account for data changes and irregularities. Furthermore, an alternative measure of a country's state of biodiversity will be employed. Given the heterogeneity of results for US and non-US banks, the baseline regression will also be estimated without US banks, to account for potential bias. The robustness checks overall confirm the presented results.

This thesis is structured as follows. Chapter 2 introduces important concepts of biodiversity and how its associated risks affect the financial sector. In Chapter 3, a literature review is conducted to provide a brief overview of related research papers and to highlight the scientific contribution of this thesis. Chapter 4 motivates the research hypothesis and introduces the econometric model. In Chapter 5, a description of the data is provided and in Chapter 6, the regression results are presented. In Chapter 7, additional analyses provide robustness checks for the presented results, and Chapter 8 concludes the thesis.

2 From Biodiversity Loss to Financial Risk

Dasgupta (2021) defines biodiversity simply as "the diversity of life" (p. 36). Described more specifically, it encompasses all ecosystems, as well as genetic variation and interaction within and between species. Additionally, many aspects of biodiversity can vary over time and therefore provide, together with genetic diversity, potential for evolution (Dudley, 2023). Valuing biodiversity is a challenge. According to Dasgupta (2021), ecosystems can be regarded as a third capital good, among human capital and produced capital. Biodiversity, however, is only a characteristic of natural capital and is therefore often disregarded when studying the value provision of nature. Dasgupta (2021) therefore defines six possibilities to value biodiversity, which blend into one another but can be distinguished. Fundamental value sources include the existence of humans, the existence of other species, as well as an intrinsic value of nature. Additionally, value can be derived from biodiversity's contribution to human health, as well as simple enjoyment of nature. These value sources have formed the basis for industries, like pharmaceuticals and ecotourism. The last source of value, which is mostly acknowledged in environmental and resource economics, are the goods and services available through nature. This source allows for an indirect valuation of biodiversity through studying industries that rely on the provision of such goods (Dasgupta, 2021). As biodiversity concerns increased in recent years, these estimations help to understand the effects global economies may face.

2.1 Risk Drivers of Biodiversity Loss

In a recent report by the WEF (2020), experts across academia, business, government, and the international community have been surveyed regarding their perceived global risks. Among other nature-related risks, biodiversity has been determined to be the 3rd most severe in the long term. This comes to no surprise, as the rate of global nature change over the past 50 years is unprecedented in human history (IPBES, 2019). According to IPBES (2019), there are five direct global change drivers, which in turn result from indirect ones. Changes in land- and sea-use, direct exploitation of organisms, climate change, pollution, and invasive species can all be traced to human interactions, like production and consumption behaviors, technological innovation, or population dynamics.

Climate change being one of the risk drivers for nature change and therefore biodiversity loss, highlights an important aspect in their relation. Conceptually, biodiversity risks are different from climate-related risks. While biodiversity risk captures "the threats to the variety of life on Earth and its consequences, [...] climate risk relates to the potential negative consequences of a change in the climate system" Giglio et al. (2023, pp. 4–5). However, there exist important similarities and interdependencies between them. While both can result from similar direct change drivers, like deforestation and pollution, they don't necessarily have to. Indirect change drivers, however, are equal for both climate change and biodiversity loss (Pörtner et al., 2021). Importantly, both risks can amplify each other and drive climate and biodiversity degradation even further. Changes in temperature or CO_2 concentration, which are direct results of climate change, influence the working balance of ecosystems (Barger et al., 2018). In turn, biodiversity and functioning ecosystems are crucial to mitigate climate change. Global forest areas are crucial carbon storages, the loss of which counteracts climate change mitigation efforts and increases climate risk (McKinley et al., 2011). However, Giglio et al. (2023) show the importance of distinguishing both risks in the context of economic implications, as they follow different temporal patterns when quantified. Therefore, this thesis focuses on biodiversity risk only.

According to Dasgupta (2021), biodiversity risk can emerge through physical, transition, or litigation risk. Physical risks refer to the impact of ecosystem degradation on economies. Transition risk is caused either directly or indirectly, through misalignment with regulations, technological progress, or changes in consumer sentiment. Litigation risks occur if legal frameworks related to natural capital are breached. Independent through which channel these risks arise, they all affect global economies. According to the Organisation for Economic Co-operation and Development

(OECD), biodiversity risks will disseminate through global economies, initially through physical risks, by affecting industries which depend highly on ecosystem services (OECD, 2023a). Value chains could be disrupted, which leads to fluctuations and eventually an increase in systematic risk. This will materialize through microeconomic impacts, like business disruptions, and eventually macroeconomic consequences like growing inflation or deviating trade flows. Such alterations are directly linked to indirect change drivers and will increase spillover effects to other industries. Consequences of biodiversity loss will therefore not only affect those industries with direct exposure but various economic sectors (van Toor et al., 2020).

2.2 Biodiversity Risk in the Financial Sector

The financial industry is one of the economic sectors which will be indirectly affected by the consequences of biodiversity loss. As there is no significant direct dependency on ecosystem services, biodiversity-related issues materialize through financial risks instead of physical risks (Hudson, 2024; WEF, 2020). Macroeconomic changes impact financial risk drivers like market risk, operational risk, and credit risk (Dasgupta, 2021). Dasgupta (2021) provides an overview of how these risk drivers are affected by biodiversity loss. Market risk may be influenced by changes in economic conditions and devaluation of assets in consequence of physical ecosystem changes. Furthermore, market shifts and regulatory policies increase transition risk and affect financial performance and asset values. Operational risks arise from physical risks, such as disruption of supply chains through nature-related changes. Additionally, transitions in consumer sentiment toward sustainable practices may result in reputational damage. Lastly, financial institutions face credit risk, caused by potential losses in corporate lending if investees default on their credit obligations. Disruption of business operations or supply chains, caused by ecosystem degradation, can negatively influence asset quality and increase non-performing assets within lending portfolios. For financial institutions, this inflates both the probability of default and loss given default. Similarly, transition risks can disrupt business models, which reduces corporate earnings of investees, who may therefore not be able to meet their credit obligations.

In recent years, the severity of potential consequences of biodiversity loss for the financial sector in general has been acknowledged (Almeida et al., 2022; CBD Secretariat, 2021a; Hudson, 2024; OECD, 2023b). However, among financial institutions, exposure to credit risk is especially high for banks, as traditional business models strongly rely on customer loans (Bolívar et al., 2023). According to an estimation of the European Central Bank (ECB), companies with high reliance on at least one ecosystem service received about 75 percent of bank loans in the Euro area (Elderson, 2023). In the USA, bank loans worth 1.7 trillion dollars and bank securities worth 300 billion dollars were identified to be exposed to nature-related risks (Deloitte, 2024). Consequently, banks are particularly prone to credit losses and potential default of clients (Ahmad & Karpuz, 2024). Therefore, this thesis focuses on banks and aims to examine the effect of biodiversity losses on their credit risks.

3 Literature Review

This thesis contributes to different strands of literature. Firstly, it expands the literature on sustainability-related risk factors and their influence on credit risk, measured in CDS. These risk factors include, for example, carbon risk (Blasberg et al., 2024; Z. Zhang & Zhao, 2022), climate-related transition risk (Costola & Vozian, 2025; Ugolini et al., 2024), and corporate sustainability performance (Christ et al., 2022). Across the board, these factors are found to significantly impact CDS pricing. Kölbel et al. (2022) additionally examine whether the disclosure of physical and transitional climate risks influences CDS spreads. For the period after the Paris Climate Agreement in 2015, they detect that the disclosure of climate-related risks can have both positive and negative effects on CDS through either higher risk perception, or lower uncertainty. This thesis examines biodiversity risk as a new factor of sustainability-related risk, adding to the existing research on nature-related credit risk determinants.

Secondly, this thesis adds to the growing literature on the influence of biodiversity risk on the finance sector. An initial approach to estimate the impact of biodiversity loss on financial systems was introduced by van Toor et al. (2020). In a joint report by De Nederlandsche Bank and PBL Netherlands Environmental Assessment Agency, they introduce an approach to measure the magnitude of financial risk from biodiversity exposure. They use the ENCORE database, which maps different business processes and their reliance on ecosystem services, to categorize financial portfolios of Dutch financial firms. They find that 36% of the examined portfolios are highly or very highly reliant on ecosystem services, showing the potentially severe impact biodiversity loss can have on the Dutch financial system. This pioneering effort has since been replicated. For instance, Calice et al. (2021) use this approach to estimate the biodiversity exposure for Brazil and find that 46% of loans allocated by banks are highly or very highly dependent on ecosystem services. Similarly, Svartzman et al. (2021) adopt the methodology for the French financial system and find that 42% of the portfolios are dependent on ecosystem services. Another implementation of this approach is the study by Calice et al. (2023), who investigate emerging markets. They find that in those markets, banks allocate around 50% of their credit portfolio to firms with high or very high reliance on ecosystem services. They also show that this ratio is higher in low-income countries and lower for high-income countries.

There has also been an evolving literature studying whether biodiversity risks are priced in financial markets. Coqueret et al. (2025) use a biodiversity measure called Corporate Biodiversity Footprint (CBF) from Iceberg Data-Lab to construct portfolios going long on firms with a low biodiversity footprint and short on firms with a high

footprint. When narrowing their analysis to sectors highly exposed to both physical and transition risk, they find that firms with a low biodiversity footprint realize higher returns. Additionally, they find that investors require a higher risk premium for assets with an excessive biodiversity footprint. This gap in expected returns has become more evident since the Kunming Declaration in 2021, which they argue with the increasing public awareness of biodiversity risks. Similarly, Garel et al. (2024) employ the CBF measure to examine stock returns by conducting an event study around the Kunming Declaration in 2021. They find that biodiversity is not priced in assets on average over their sample. However, after the external shock, firms experienced decreasing stock returns, which were more severe for firms with a high CBF. Their result also demonstrates investors' growing awareness of potential risks related to biodiversity since the Kunming Declaration and consequently a pricing of those risk factors in stock prices. Créti et al. (2024) follow a different approach in assessing the pricing of biodiversity in stock markets. They construct a new index from a "bio-capacity" and "bio-footprint" measure to relate a country's biodiversity richness to the damage inflicted upon it. On a firm level, they find a significant influence of physical biodiversity risks on profitability and stock returns. Additionally, Naffa and Czupy (2024) also find a biodiversity risk premium in stock markets. For their analysis, they use the MSCI's ESG scores to examine biodiversity performance on a firm level.

These approaches all rely on different data sources for biodiversity risk, all with their own shortcomings. The approach pioneered by van Toor et al. (2020) relies on the ENCORE database, which underestimates biodiversity reliance, as only first-order dependencies are considered (Calice et al., 2021). The CBF relies in large parts on estimates and sector averages to provide firm-level characteristics and is only available on an annual frequency (Garel et al., 2024). The measure introduced by Créti et al. (2024) only represents a country-level biodiversity resilience statistic, which does not capture the time-varying trends and developments contained in the measure of Giglio et al. (2023). MSCI's ESG score for biodiversity and land use mainly evaluates the management of the included risks and does not provide a criterion for the biodiversity exposure of external stakeholders (MSCI ESG Research LLC, 2023). When analyzing banks, such risks are important to assess their overall credit risk.

Giglio et al. (2023) recognize the need for a suitable quantification of both physical and transition risks. In their paper they, introduce text-based measures to quantify firm and industry-level biodiversity risks in the USA. Additionally, they construct a biodiversity news index, which quantifies the presence of biodiversity topics in New York Times articles. This indicator provides a comprehensive high-level risk measure, which mirrors overarching risks faced on a global scale. Using their firm-level metric, they find that biodiversity risks are already priced in equity markets. Based on this work, many papers investigating the effects of biodiversity risks on finance followed, which employ these measures. Kalhoro and Kyaw (2024) concudct an event study to show that around worldwide discussions on biodiversity risks, markets respond by adjusting firm valuations, depending on the risk exposure of industries and firms. Liang et al. (2024) find that a higher biodiversity risk exposure increases the crash risk of stocks. Ahmad and Karpuz (2024) show that firms with higher biodiversity risk vulnerability tend to gather higher cash reserves. Bach et al. (2024) analyze firms' performance and find that higher biodiversity risk negatively impacts growth and profitability. This thesis adds to this emerging body of biodiversity finance literature, which employs the risk measure by Giglio et al. (2023), by examining the effect of biodiversity risk on banks' CDS spreads.

Lastly, this thesis also contributes and extends initial approaches to examine the relationship between biodiversity risk and credit risk measured in CDS. Hoepner et al. (2023) investigate the impact of biodiversity, water scarcity, and pollution prevention on the financing conditions of infrastructure companies. They find that firms which manage these risks experience a significant decreasing effect on CDS slopes. They follow that the results suggest that these risks are perceived as long-term issues. Giglio et al. (2024) develop and empirically test a model to show how biodiversity influences economic output. In their analysis, they use the biodiversity news index developed by Giglio et al. (2023) to examine its effect on CDS spreads of government-issued bonds. They find that CDS spreads rise when negative news about biodiversity emerges. Furthermore, they show that this effect is weaker in countries with a superior state of biodiversity and a higher share of natural capital of overall wealth.

This thesis extends the work by Hoepner et al. (2023) and Giglio et al. (2024) by adapting their approach to the banking industry. In recent analysis on biodiversity finance, these firms are often excluded, leaving the effects of biodiversity on the banking industry relatively unclear (Créti et al., 2024; Liang et al., 2024). This thesis aims to close this gap.

4 Hypothesis Development & Econometric Model

As shown in Chapter 2, the effects of biodiversity loss affect financial institutions through several channels. Importantly, banks are sensitive to increasing credit risk because of their reliance on customer loans. This thesis examines whether biodiversity risk is significantly affecting credit risk in the banking industry.

As a measure of credit risk, CDS are used in this thesis. CDS are credit derivatives which serve as insurance in case of default of the underlying (Ericsson et al., 2009). In a CDS contract, the buyer pays a periodic premium called the CDS spread, which is denoted in basis points (bp) of its notional value (Zhu, 2006).[1] If the spread amount is denoted at 100 bp (1.0%), a buyer would need to pay a premium of $100,000 to insure debt with notional value of $10 million

[1] The spread and price of a CDS contract refer both to the premium paid by the buyer of the contract and are both used interchangeably throughout this thesis.

against default (Hasan et al., 2016). These premium payments are made either until maturity of the contract or the occurrence of a credit event (Blanco et al., 2005). Credit events include bankruptcy, obligation acceleration, obligation default, failure to pay, repudiation or moratorium, and restructuring (Zhu, 2006). In return for the premium payments, the seller of a CDS contract agrees to compensate the buyer for the losses incurred in case of a credit event (Ericsson et al., 2009). Therefore, CDS are a way to trade and also short credit risk, at a pre-defined price (Blanco et al., 2005). A bank's CDS spread therefore provides a measure of how its credit risk is perceived by the market (Wang et al., 2019). Compared to other credit risk measures, CDS provide distinct advantages. First, CDS contracts are homogeneous and standardized, which enables uniform usage in cross-country samples (Hasan et al., 2016). Additionally, the CDS markets respond more quickly to changes in credit risk and contain exclusive credit information, compared to bond yields (Blanco et al., 2005; Lee et al., 2018). CDS are also less affected by non-default components of risk, compared to bond spreads, which are more likely to be affected by contractual agreements and liquidity factors (Longstaff et al., 2005; B. Y. Zhang et al., 2009).

As a measure of biodiversity risk, this thesis employs the biodiversity news index by Giglio et al. (2023). In their paper, they recognize the need for adequate measurements and quantification of biodiversity risk. As a response to this issue, they develop a news-based biodiversity measure. The index is constructed by implementing a biodiversity dictionary, which includes biodiversity-related terms. New York Times articles are then analyzed to detect sentences which include at least one of the identified terms. To account for both negative and positive news about biodiversity risks, they determine the sentiment of each sentence by using the language model BERT. The number of positive news are then subtracted from the number of negative news, resulting in an index, where a higher value relates to more negative news. Focusing on newspaper articles has the advantage that current developments are quickly captured and asset markets reflect such news immediately (Giglio et al., 2024). But this process might cause the following shortcoming. Given only New York Times articles are considered, the index might over-represent biodiversity related topics, which are relevant to the USA but not on a global scale. Therefore, the question arises, whether the index is representative of global biodiversity concerns. However, Giglio et al. (2024) successfully apply the index to a global sample to investigate effects of biodiversity risks on sovereign CDS. Furthermore, this thesis addresses the potential issue by performing a robustness check, which excludes US banks to show that the obtained results are equally found for non-US banks and are therefore not driven by selection bias.

If credit risk in the banking sector is driven by biodiversity risks, this should be reflected in CDS pricing. During periods with negative news about biodiversity, the CDS spreads should react accordingly, as they are a measure of the market's bank credit risk perception (Wang et al., 2019). Therefore, the first tested hypothesis is the following:

> *Hypothesis 1. Bank's CDS spreads are positively related to biodiversity news and rise when negative news emerges.*

Following the Kunming Declaration in October 2021, biodiversity risks have received increasing global attention (Garel et al., 2024). Participating countries especially focused on the need to rethink the role of the financial sector regarding their investments and disclosure of biodiversity contributions (CBD Secretariat, 2021a). Similar to the Paris Agreement in 2015, the signing of the Kunming Declaration may encourage national and subnational policies, which increases transition risk (Almeida et al., 2022). Research has shown that impacts of biodiversity on the financial sector have been more severe after the Kunming Declaration (Garel et al., 2024; Kalhoro & Kyaw, 2024). Similarly, the increased perception of biodiversity risk following the Kunming Declaration may also influence the valuation of banks' credit risk. Therefore, a second hypothesis is tested:

> *Hypothesis 2. After the Kunming Declaration in 2021, the positive relationship between negative biodiversity news and banks' CDS spreads is more severe, compared to the pre-Kunming period.*

Giglio et al. (2024) find that the magnitude of biodiversity risk effects on CDS prices differs across countries. Characteristics like the state of biodiversity and the dependency on natural capital seem to influence how CDS react to negative news on a country level. For banks, the exposure to biodiversity risk is in part channeled through their credit portfolios (Hudson, 2024). Firms within these portfolios may experience varying effects, depending on their geological environment. Consequently, banks may experience heterogeneous impacts of increases in biodiversity risk, based on their location. Therefore, the following hypothesis is tested:

> *Hypothesis 3. The positive relationship of biodiversity news and bank's CDS spreads is stronger for banks located in countries with an inferior state of biodiversity.*

Transition risks affect banks through changes in market conditions. To mitigate this risk, they need to be open to adapting technological innovations and changing business processes (Dasgupta, 2021). Disclosing such intentions may be well received by market participants, which might influence CDS pricing. Kölbel et al. (2022) find that CDS markets respond to the disclosure of physical and transitional climate risks. For banks, there exist global initiatives, which aim to drive the adoption of sustainable practices and international agreements (Hudson, 2024). Adopting such initiatives may pose similar effects for banks, as those shown by Kölbel et al. (2022). Therefore, the following hypothesis is formulated:

> *Hypothesis 4. Banks which proactively acknowledge biodiversity risks experience lower impacts of such on their CDS spreads.*

The CBD consists of 196 member states, more than 100 of which have committed to implement the global biodiversity framework and engage in conservation efforts until 2030 (CBD, 2025; Garel et al., 2024). The USA is the only UN member to sign but not ratify the treaty to this day (Einhorn, 2025). Although there have been efforts towards biodiversity conservation, its influence on the agreements of the CBD is therefore limited. With the adoption of the global biodiversity framework in the context of COP 15, participating countries agreed to implement reporting obligations. These commitments are not legally binding, but have accelerating impacts on biodiversity-related regulation and litigation (Garel et al., 2024). Due to the lack of commitment of the USA to the global framework, the US economy and their banks might face lower transition risk due to the Kunming Declaration, which should be reflected in the relationship of biodiversity news and CDS price changes. Therefore, the following hypothesis is tested:

> *Hypothesis 5. Banks located in the USA experience lower effects of biodiversity news on CDS spreads after the Kunming Declaration in 2021.*

By adding country and bank characteristics to the analysis, this thesis aims to not only estimate the direct relationship between CDS and biodiversity risk but also to identify underlying drivers, which change the magnitude of those effects.

The baseline empirical design relies on panel regression to examine the relationship between CDS and biodiversity risks across entities and time. The regression is a multivariate OLS regression with different sets of fixed effects and clustered standard errors. It is estimated as follows:

$$CDS_{i,t} = \alpha + \beta_1 Biodiversity_t + \beta_2 X_{i,t} + \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t} \quad (1)$$

The variable of interest $CDS_{i,t}$ denotes quarterly CDS returns for bank i at time t. The main explanatory variable $Biodiversity_t$ denotes the quarterly average of the biodiversity index by Giglio et al. (2023) at time t, which is invariant across banks and only varies over time. $X_{i,t}$ contains various bank and country characteristics as control variables. Bank characteristics include the leverage ratio, asset quality, bank size, funding stability, cost efficiency, and sensitivity to market risk. Furthermore, bank's equity return volatility is included in the analysis. For country characteristics, the stock index return and volatility, the slope of the yield curve, as well as the sovereign credit rating, are used. As fixed effects, time fixed effects γ_t, bank fixed effects μ_i and country fixed effects φ_j are employed. Lastly, α denotes the intercept, and $\epsilon_{i,t}$ the error term.

5 Data

Chapter 5 presents the data used in this thesis. First, the variables and their sources are described. Following, the unit root test is undertaken to account for non-stationarity. Lastly, the summary statistics of the sample are presented.

5.1 Variable Description

To create panel data for the banks in the analysis, CDS and the biodiversity index are merged with bank- and country-level control variables. Table 1 shows an overview of the variables included in the sample.

Data for CDS is available through Refinitiv Datastream. This database provides coverage of over 112,000 active CDS and further financial markets data (Refinitiv Datastream, n.d.). First, the available CDS are filtered to exclude inactive data series. Afterwards, all CDS of firms which are not banks, or which refer to sovereign debt, are excluded. CDS contracts can have different maturities as well as different restructuring clauses, which define under which circumstances changes to the debt obligations are considered a credit event (IHS Markit, 2021). Following Jorion and Zhang (2007), CDS with a 5-year maturity are used, as they are the most liquid and constitute the majority of contracts in CDS markets. Additionally, only those contracts are included, which cover senior unsecured debt and are denoted in US dollars. In contrast to Jorion and Zhang (2007), CDS data is not filtered regarding its restructuring clause. Following the financial crisis in 2009, the International Swaps and Derivatives Association (ISDA) redefined contractual characteristics, including changes to restructuring and credit events. Main protocols include the Big-Bang protocol from 2009 and ISDA 2014 definition modifications (Boyarchenko et al., 2019). Concluding these changes, different economic regions have adopted contracts with different restructuring clauses as their standard (Boyarchenko et al., 2019; IHS Markit, 2021). Contracts with full restructuring (CR) are mainly traded in Asia, contracts with modified-modified restructuring (MM) in Europe, and contracts with no restructuring (XR) in North America (IHS Markit, 2021). Furthermore, there are contracts with modified restructuring (MR), for which no specific region is mentioned by IHS Markit (2021). When constructing the sample, they are found to be implemented in CDS contracts for Australian banks and included accordingly. To account for these contractual changes and provide a global sample of banks, CDS contracts are therefore considered with all restructuring clauses. In case there exists CDS data with different restructuring clauses for the same bank, the contract is chosen, which is predominantly traded in the bank's country. In Appendix A.1, a list of CDS contracts and their restructuring clause is provided for the included banks. The result is a sample of CDS contracts for 75 banks. After excluding banks for which control variable data is not available, the total sample consists of 39 banks. CDS data considered in this thesis starts in the year 2015 to provide contractual conformity, as the 2014 definition changes to contractual terms by the ISDA came into effect in September 2014 (Boyarchenko et al., 2019).

Table 1: Variable Description

This table shows a description of the variables used in the sample, as well as the source and expected sign of the regression coefficient.

Variable	Description	Source	Expected Sign of Coefficient
CDS	Quarterly return of bank's CDS spreads	Refinitiv Datastream	
Biodiversity	Quarterly mean of the biodiversity index by Giglio et al. (2023)	Biodiversityrisk.org	+
Leverage	Ratio of total liabilities to the sum of total liabilities and total equity	Fitch Ratings	+
Asset Quality	Ratio of loan loss provisions to total loans	Fitch Ratings	+
Bank Size	Natural logarithm of total assets	Fitch Ratings	-
Funding Stability	Ratio of total deposits to total liabilities	Fitch Ratings	-
Cost Efficiency	Ratio of operating expenses to total revenue	Fitch Ratings	+
Sensitivity to Market Risk	Ratio of interest expenses to total liabilities	Fitch Ratings	+
Bank Volatility	Quarterly standard deviation of daily bank equity returns	Refinitiv Datastream	+
Stock Market Index Return	Quarterly return of a country's stock market index	Refinitiv Datastream	-
Stock Market Index Volatility	Quarterly standard deviation of daily stock index returns	Refinitiv Datastream	+
Yield Curve Slope	10-year government bond yield less 2-year government bond yield	Refinitiv Datastream	-
Sovereign Credit Rating	Index ranging from 1 (Fitch Ratings CCC rating) to 17 (Fitch Ratings AAA rating)	Fitch Ratings	-

Figure 1 shows the average daily CDS spread of the banks in the sample over the period between 2015 and 2023. The average spread of CDS in the sample was fluctuating between 60 and 180 basis points, with several periods of sharp increases. During the period between 2015 and early 2016, the average spread was rising, until it reached a maximum of about 176 basis points in February 2016. Additionally, spreads experienced a drastic rise from about 70 basis points to 145 basis points in early 2020. Furthermore, spreads were increasing drastically in the first months of 2022. When investigating these periods of sharply increasing CDS spreads in the past years, it becomes evident that CDS react to economic and global events. In early 2016, the price of oil had experienced the low point of a period of plunging prices from mid-2014 onwards (Stocker et al., 2018). This was followed by a slowdown in global growth, tighter fiscal policies, and increased credit risk for banks in the energy sector (Stocker et al., 2018; World Bank Group, 2016). Furthermore, the global COVID-19 pandemic, officially declared in early 2020, has damaged human lives, global economies, and credit conditions (Yin et al., 2022). Additionally, the Russian invasion of Ukraine in early 2022 imposed challenges on global supply chains and energy infrastructure, while also driving inflation, especially in the Euro area (Cui et al., 2023; Ferreira et al., 2025). While the sample in this thesis only includes a cross-section of global banks, the impacts of such events are still represented as spikes in the average CDS spreads. To account for such external shocks and to isolate their impact, time fixed effects are employed in the regression. However, the period of the COVID-19 pandemic will need to be handled with additional care. Compared to other times of crisis, the year 2020 has seen enormous spikes in economic uncertainty, similar in magnitude to the Great Depression of 1929-1933 (Baker et al., 2020). Through radical governmental measures, which were introduced to halt the spreading pandemic, firms experienced severe consequences on demand, revenue, and profitability, among others. The banking sector was among the hardest-hit sectors and reacted to increased uncertainty with a spike in CDS prices, as shown in Figure 1 (Apergis et al., 2022). Additionally, rising infection numbers induced these price increases (Hasan et al., 2023). Therefore, the pandemic and especially the year 2020 needs to be regarded as an exceptional period of crisis, in which CDS prices were significantly driven by pandemic metrics like confirmed cases and deaths (Apergis et al., 2022). Due to these factors, results of credit risk analyses during this period may be distorted. As a consequence, this thesis follows the approach by Blasberg et al. (2024), who exclude the year 2020 from their sample when examining the effect of carbon risk on CDS prices.[2]

[2] In Appendix A.5, regression results with the full sample are disclosed.

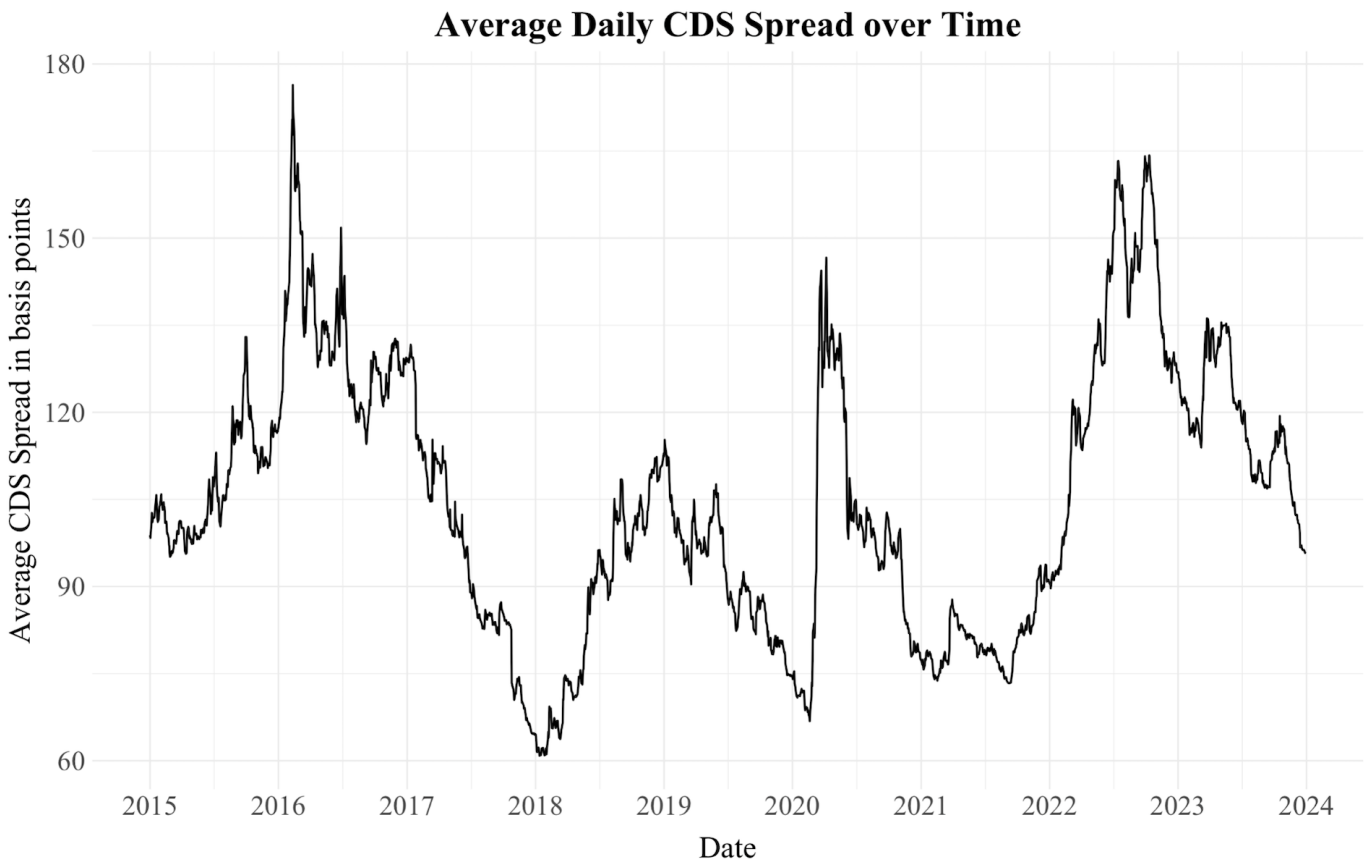

Figure 1: Average Daily CDS Spread of Banks in the Sample between 2015 and 2023.

This figure shows the daily mean of CDS spreads of the banks in the sample over the sample period between 2015 and 2023. The x-axis depicts the time and the y-axis the average CDS spread, which is depicted in basis points.

The data for the biodiversity index by Giglio et al. (2023) can be downloaded from *biodiversityrisk.org*. Observations of the index start in the year 2000 and end in 2023. For the period between 2015 and 2023, the average of daily index values is calculated for each quarter. A higher value of the biodiversity index is associated with a higher amount of bad news in the corresponding period. The biodiversity data is then merged with the CDS data for each bank. In Figure 2, the solid line shows the quarterly average of the biodiversity index. The dashed line depicts a fitted linear regression line, which shows the time trend of the biodiversity index. The graph shows the high fluctuation of the index over the sample period. Values range between 0.119 in Q3 of 2016 and 0.554 in Q3 2019 and Q3 2022. Across the sample period, the average biodiversity index is positive, which shows that news about biodiversity had a predominantly negative sentiment. Additionally, the slope of the fitted regression line is positive, which suggests that the biodiversity index takes on higher values in later periods of the sample.

Additionally, control variables are included, which have been shown to have significant effects on CDS of global banks (Drago et al., 2017; Hasan et al., 2016). First, balance sheet variables are included, which are available through Fitch Ratings. The leverage ratio is calculated as total liabilities divided by the sum of total liabilities and total equity. An increase in a bank's leverage ratio is expected to have increasing effects on CDS price changes (Hasan et al., 2016). Asset quality is a measure of quality and trends of a bank's major assets. It evaluates the management of credit risks, like the quality of loan underwriting or handling of non-performing assets (Hasan et al., 2016). Following Hasan et al. (2016), the loan loss provision ratio is considered as a measure for asset quality, which is defined as the ratio of loan loss provisions to total loans. An increase in the loan loss provision ratio is signaling lower asset quality, which is expected to increase CDS spreads. Bank size is included, as larger banks may be able to better diversify risk and may be deemed as too big to fail (Drago et al., 2017). As a measure for bank size, the natural logarithm of total assets is included. An increase in bank size is expected to decrease CDS spreads. Funding stability is included and measured by the ratio of deposits to total liabilities, as deposits are considered to be a stable source of funding (Drago et al., 2017). An increase in funding stability is therefore expected to be related to lower

While the baseline results do change, the overall findings of this thesis are supported. Please refer to the results in the appendix for more details.

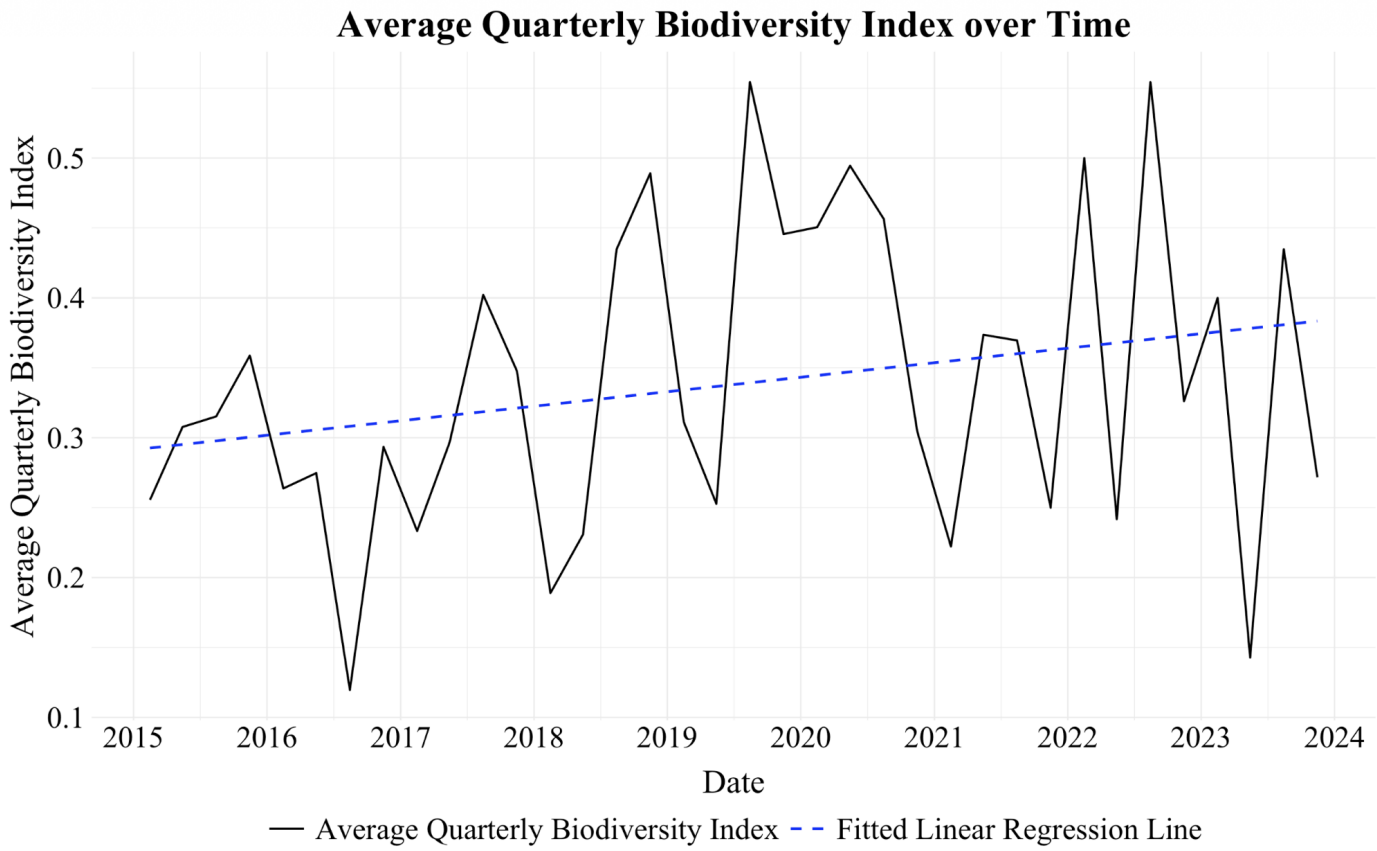

Figure 2: Average Quarterly Biodiversity Index between 2015 and 2023.

This figure shows the quarterly mean of the biodiversity index over the sample period between 2015 and 2023. The x-axis depicts the time and the y-axis the average biodiversity index. The dashed line depicts a fitted linear regression line, which shows the time trend of the average biodiversity index. The data points are displayed on the median day of each quarter.

CDS spreads. Additionally, cost efficiency is included as a measure of management quality. High-quality management helps to efficiently identify and navigate a bank's risks and ensure operation with low credit risk (Hasan et al., 2016). Cost efficiency is calculated as the ratio of operating expenses to total revenues, where a higher value shows less efficiency. Therefore, CDS spreads are expected to increase if the ratio increases. As the last balance sheet variable, sensitivity to market risk is included, which is the measure of sensitivity to shifts in interest rates for all loans and deposits. To account for this interest rate risk, cost of fund is included as a control variable, which is the ratio of interest expense to total liabilities. Banks with a higher cost of fund are considered to be more prone to changes in market conditions (Hasan et al., 2016). Increasing sensitivity to market risk is therefore expected to be associated with higher CDS spreads. Balance sheet variables are available for the 39 banks on a half-yearly basis. As the sample is constructed on a quarterly basis, the balance sheet variables are interpolated to match the quarters within a half-year period with the same value.

In addition, several market variables are included as control variables as well. First, a bank's equity returns are used to calculate its equity return volatility, which has been shown to have a significant impact on CDS prices (Hasan et al., 2016). It is calculated by taking the standard deviation of daily equity returns within a quarter. Higher equity volatility is associated with higher risk and should therefore increase CDS spreads (Tang & Yan, 2010). Additionally, stock index return is included to capture a country's business climate (Drago et al., 2017). Returns are calculated as the quarterly return of a country's stock index. Higher returns indicate a better business climate, which is expected to decrease CDS prices. Furthermore, the stock index return volatility is included, which is derived similarly to equity return volatility. Daily stock index returns are calculated, before their standard deviation within each quarter is derived. Similarly to equity return volatility, higher index volatility indicates greater economic uncertainty and default risk (Tang & Yan, 2010). A country's yield curve slope is calculated as the difference between the return on 10- and 2-year government bonds and is depicted in percent or basis points. It is also included as a control variable, as the slope is considered to be an economic indicator, where a higher yield curve slope suggests a better economic outlook (Hasan et al., 2016). Hence, a higher yield curve slope is expected to decrease CDS spreads. Additionally, sovereign credit ratings are included in the analysis,

Table 2: Panel Unit Root Test

This table shows the test results of the Fisher-ADF unit root tests by Choi (2001). Depicted are the dependent and explanatory variables as absolute values and as period change. Two tests are undertaken, the inverse normal test in the first two columns and the logit test in the last two columns.

Panel unit root test				
	Inverse normal		Logit	
	Statistic	P-value	Statistic	P-value
CDS	0.29	0.62	0.45	0.67
Δ CDS	-30.32	0.00	-52.11	0.00
Biodiversity	1.20	0.89	1.06	0.86
Biodiversity (standardized)	-37.50	0.00	-71.86	0.00
Leverage	4.37	1.00	4.30	1.00
Δ Leverage	-13.99	0.00	-18.60	0.00
Asset Quality	-1.05	0.15	-1.82	0.04
Δ Asset Quality	-12.98	0.00	-15.51	0.00
Bank Size	7.62	1.00	7.93	1.00
Δ Bank Size	-11.65	0.00	-14.13	0.00
Funding Stability	8.68	1.00	8.86	1.00
Δ Funding Stability	-15.23	0.00	-20.93	0.00
Cost Efficiency	-0.09	0.46	-0.66	0.25
Δ Cost Efficiency	-16.61	0.00	-25.81	0.00
Sensitivity to Market Risk	6.86	1.00	7.16	1.00
Δ Sensitivity to Market Risk	-7.62	0.00	-8.00	0.00
Bank Volatility	-0.52	0.70	1.22	0.89
Δ Bank Volatility	-41.12	0.00	-86.54	0.00
Stock Market Index Return	7.51	1.00	7.43	1.00
Δ Stock Market Index Return	-33.05	0.00	-58.42	0.00
Stock Market Index Volatility	-2.44	0.01	-2.20	0.01
Yield Curve Slope	-7.64	0.00	-7.61	0.00

where the credit rating assigned by Fitch Ratings is matched with a numeric scale. A full description of the rating scale is shown in Appendix A.2. As credit ratings are a depiction of a country's leverage and credit risk, a higher rating should be associated with a decrease in CDS prices. After accumulating all balance sheet and market control variables, they are merged with a bank's CDS data and the biodiversity measure. Country-level data is matched with each bank according to the location of its headquarters. All continuous variables except the biodiversity index and credit ratings are winsorized at the 1% and 99% levels to account for the influence of outliers (Garel et al., 2024). In total, the sample consists of 1,248 observations for 39 banks between 2015 and 2023, excluding the year 2020.

5.2 Unit Root Test

The variables of the sample may not be included as their absolute values, because of potential non-stationarity, where variables follow time-persistent trends. Stationarity, however, is important in the sense that estimated relationships should be stable over time when applying OLS regression (Wooldridge, 2020). Following Corò et al. (2013), the Fisher-ADF panel unit root test by Choi (2001) is used to examine the variables in the sample for unit roots. If they contain a unit root, it suggests that there exist time-persistent trends, which indicates non-stationarity. There are two unit root tests introduced by Choi (2001), called inverse normal and logit test. Both test that all entities contain a unit root as the null hypothesis against the alternative hypothesis that some do not (Corò et al., 2013). Table 2 shows the results of both unit root tests applied to the sample once in absolute values and once depicted as their period change. Credit ratings are not tested, as there exist entities without variation where the application of the unit root test is not feasible (Wooldridge, 2020). For p-values below 0.05, the

null hypothesis is rejected and the alternative hypothesis accepted, which implies stationarity. The results show clear indications that the null hypothesis is not rejected for all variables except stock market index volatility and yield curve slope. The other variables reveal non-stationarity and need to be transformed. All variables except biodiversity, stock index volatility, the yield curve slope, and credit rating are depicted as the period change in percent. For asset quality, the p-value is below 0.05 in the case of the logit test but above for the inverse normal test. As this result does not provide a clear implication, the following approach is taken. To ensure uniformity across balance sheet variables, asset quality is also included in the analysis as its period change. In robustness checks, it will be shown that including asset quality in absolute values provides similar results. The biodiversity

Table 3: Summary Statistics

This table shows the summary statistics for the variables used in the baseline regression. The sample includes a total of 1,248 quarterly observations for 39 banks between 2015 and 2023, excluding the year 2020. All variables except biodiversity and sovereign credit rating are winsorized at the 1% and 99% level.

Summary Statistics						
Baseline Sample *N = 1,248*	Mean	St. Dev.	Min	25th percentile	75th percentile	Max
Δ CDS (%)	1.779	22.787	-41.954	-13.212	10.383	80.107
Biodiversity (standardized)	0.000	1.000	-1.916	-0.692	0.680	2.100
Δ Leverage (%)	0.019	0.694	-1.928	-0.334	0.357	2.537
Δ Asset Quality (%)	-2.395	11.475	-31.456	-8.870	2.882	42.559
Δ Bank Size (%)	0.014	0.219	-0.692	-0.095	0.135	0.610
Δ Funding Stability (%)	0.960	5.076	-14.488	-1.498	3.071	18.731
Δ Cost Efficiency (%)	1.011	17.854	-44.007	-8.640	8.269	63.988
Δ Market Sensitivity (%)	19.196	53.077	-36.578	-5.614	21.753	329.719
Δ Bank Volatility (%)	2.785	33.423	-48.505	-19.880	20.427	113.394
Δ Index Return (%)	1.812	7.215	-15.412	-2.588	6.750	21.060
Index Volatility (%)	0.958	0.374	0.386	0.693	1.143	3.234
Yield Curve Slope	0.574	0.687	-2.058	0.216	1.079	2.018
Δ Sovereign Credit Rating	-0.020	0.189	-2.000	0.000	0.000	1.000

variable is standardized to remove non-stationarity, following Giglio et al. (2024). Credit Ratings are included as the absolute change in the numerical rating system (Corò et al., 2013). Stock market volatility, as well as yield curve slope, are already stationary and included as their absolute values. As shown in Table 2, all p-values are below 0.05 after the variable transformation. The null hypothesis can therefore be rejected, which provides evidence for the variables to be stationary.

5.3 Summary Statistics

Table 3 shows the summary statistics for the variables in the baseline sample, as they are used in the regression. The average CDS change per quarter is 1.779%, with a standard deviation of 22.787, which shows the large variation between periods where CDS increased and decreased within the sample. The biodiversity index is standardized to have a mean of zero and a standard deviation of one. For the balance sheet variables, the average period change was 0.019% for leverage, -2.395% for asset quality, 0.014% for bank size, 0.960% for funding stability, 1.011% for cost efficiency, and 19.196 % for sensitivity to market risk. The average change in volatility of daily bank equity returns is 2.785%. Changes in quarterly stock index returns were on average 1.812 % and their daily return volatility 0.958%. The yield curve slope was at 0.574 on average. Lastly, the average change in sovereign credit ratings is slightly negative, showing that there has been a surplus of credit rating downgrades during the sample period.

In Table 4, the pairwise correlation matrix for the sample is shown. Correlations are relatively low across the board. With the dependent variable CDS, correlations are in line with the expected directions for all variables, except for cost efficiency, yield curve slope, and sovereign credit rating. Correlation between equity and stock index variables is sometimes of concern (see Corò et al. (2013)), but not for the specification of this sample, as all correlations between bank return volatility, stock index return, and stock index volatility are not concerning.

6 Results

In Chapter 6, the results of the regression are presented. Chapter 6.1 presents the findings of the baseline regression, by introducing the different sets of variables step by step. First, the biodiversity measure is introduced, afterwards the balance sheet variables and lastly the market variables. Chapter 6.2 extends the baseline regression and introduces the Kunming Declaration as an external shock, to further examine the relationship between biodiversity news and CDS spreads. Chapter 6.3 examines the interaction effect of a country's state of biodiversity. In Chapter 6.4, the impact of adopting biodiversity-related bank initiatives on the relationship between CDS changes and biodiversity is tested. Lastly, Chapter 6.5 tests the relationship for banks in the USA.

6.1 Impact of Biodiversity News on Banks' Credit Risk

In Chapter 6.1, the influence of biodiversity news on CDS spread changes is examined to test *Hypothesis 1*. To analyze this relationship, a multivariate panel regression is con-

ducted, initially using only the biodiversity measure. The regression formula is as follows:

$$\Delta CDS_{i,t} = \alpha + \beta_1 Biodiversity_t + \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t} \quad (2)$$

The baseline regression is split into four separate parts. The analysis is conducted once without fixed effects and afterwards with time fixed effects, time and bank fixed effects, as well as time and country fixed effects. As the biodiversity measure only varies over time but not across banks, time fixed effects at the quarterly level would fully absorb its influence. Therefore, a higher frequency is chosen and time fixed effects are included on a half-yearly basis. This approach is in line with Giglio et al. (2024), who encounter the same issue and also resort to time fixed effects at a higher frequency. To account for autocorrelation, bank-level clustered standard errors are included. Results of the regression are shown in Table 5.[3] Column one shows the regression results without fixed effects. The coefficient of *Biodiversity* is positive, as expected in *Hypothesis 1*. The relationship is significant at the 1% significance level. An increase of one standard deviation in the biodiversity measure is associated with an increase in CDS spread changes by 1.946 percentage points with a standard error of 0.507. When including time fixed effects, the coefficient and standard error of *Biodiversity* increase to 4.231 and 0.795, respectively, while still being significant at the 1% significance level. With time fixed effects, an increase of one standard deviation of the biodiversity variable is now related to an increase in CDS price changes of 4.231 percentage points. In terms of model fit, time fixed effects increase the adjusted R^2 from 0.6% to 31.9% compared to column one. Including bank or country fixed effects does not change the results, as shown in columns three and four. The coefficient of the biodiversity measure stays robust and is significant across the board. Using additional fixed effects decreases the adjusted R^2 slightly. This is likely due to additional degrees of freedom being used by the fixed effects, without adding information to the model, which decreases the adjusted R^2 (Wooldridge, 2020).

Next, the balance sheet control variables are included in the analysis. The regression formula now changes to the following:

$$\begin{aligned}\Delta CDS_{i,t} = \alpha &+ \beta_1 Biodiversity_t + \beta_2 \mathrm{LEV}_{i,t} + \beta_3 \mathrm{AQ}_{i,t} \\ &+ \beta_4 \mathrm{Size}_{i,t} + \beta_5 \mathrm{FS}_{i,t} + \beta_6 \mathrm{CE}_{i,t} \\ &+ \beta_7 \mathrm{MS}_{i,t} + \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t}\end{aligned} \quad (3)$$

$\Delta\mathrm{LEV}_{i,t}$ is the leverage ratio, $\Delta\mathrm{AQ}_{i,t}$ the asset quality, $\Delta\mathrm{Size}_{i,t}$ the bank size, $\Delta\mathrm{FS}_{i,t}$ the funding stability, $\Delta\mathrm{CE}_{i,t}$ the cost efficiency, and $\Delta\mathrm{MS}_{i,t}$ the market sensitivity. These variables are all included as their period change, as indicated by the preceding delta sign. For this regression, two different analyses are undertaken. First, the sample includes the balance sheet variables of the corresponding time periods without any adjustments. For the second analysis, a time lag is introduced for the balance sheet variables, following the argumentation of Drago et al. (2017) that information contained in such variables is not immediately available to the public. Therefore, a half-year time lag is introduced, which corresponds to the frequency of data availability. As banks conduct their operations during a half-year period, the results are made public in the following period. Since CDS prices are based on market perception, the information obtained from balance sheet variables may be reflected in price changes in the period after they occur. The summary statistics and correlation matrix presented in Chapter 5.3 cover the sample without time lag. The statistics only differ slightly for the balance sheet variables with time lag and are therefore disclosed in Appendix A.3 and A.4. For the regression with time lag, the econometric model changes to the following:

$$\begin{aligned}\Delta CDS_{i,t} = \alpha &+ \beta_1 Biodiversity_t + \beta_2 \mathrm{LEV}_{i,t-1} + \beta_3 \mathrm{AQ}_{i,t-1} \\ &+ \beta_4 \mathrm{Size}_{i,t-1} + \beta_5 \mathrm{FS}_{i,t-1} + \beta_6 \mathrm{CE}_{i,t-1} \\ &+ \beta_7 \mathrm{MS}_{i,t-1} + \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t}\end{aligned} \quad (4)$$

Compared to formula (3), the balance sheet variables now capture the values of the previous time period $t-1$. Table 6 contains the results for the sample without time lag. Columns one to six show the stepwise introduction of the balance sheet variables using time fixed effects. Columns seven to nine show the results of the regression with all balance sheet variables and fixed effects. All regressions use clustered standard errors at the bank level to account for autocorrelation. The coefficient of *Biodiversity* is significant at the 1% level across all regressions. An increase of one standard deviation is associated with an increase in CDS price changes of 4.231 percentage points across the board, which is equal to the regression without balance sheet variables. Standard errors also remain stable across regressions and are comparable to the results presented in Table 5. All balance sheet variables are statistically insignificant across the board, which is unexpected, but left to be confirmed in the following baseline regression with all control variables. When comparing to the results in Table 5, the introduction of balance sheet variables does not increase the adjusted R^2 and hence the model fit. Furthermore, all coefficients, are in line with their expected sign, except ΔFS. However, since the coefficient is insignificant and the standard errors is relatively high, there is no evidence for a relationship in either direction. Table 7 shows the results of the regression with time-lagged balance sheet variables. The coefficient of biodiversity is unchanged compared to the results shown in Table 6. While all balance sheet variables are not significant, the introduction of the time lag changes the coefficients. Comparing columns seven to nine in Table 6 and Table 7 reveals the following changes. The sign of coefficient of the ΔFS variable now depicts the expected effect compared to the results without time lag. How-

[3] For the regressions throughout this thesis, significance levels at the 1%, 5%, and 10% levels are chosen, which are common in CDS literature. See, e.g. Corò et al. (2013), Giglio et al. (2024), Hasan et al. (2016), Hoepner et al. (2023), and Ugolini et al. (2024), and Z. Zhang and Zhao (2022) as references.

Table 4: Correlation Matrix

This table shows the pooled Pearson correlation for the variables used in the baseline regression. The sample includes a total of 1,248 quarterly observations for 39 banks between 2015 and 2023, excluding the year 2020. Depicted are the explanatory variables as used in the regression, winsorized at the 1% and 99% level.

Correlation Matrix														
Baseline Sample (N = 1,248)		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	Δ CDS (%)	1.000												
(2)	Biodiversity (standardized)	0.085	1.000											
(3)	Δ Leverage (%)	0.107	-0.135	1.000										
(4)	Δ Asset Quality (%)	0.027	0.066	-0.083	1.000									
(5)	Δ Bank Size (%)	-0.134	-0.060	0.216	-0.104	1.000								
(6)	Δ Funding Stability (%)	-0.024	0.055	-0.218	-0.143	-0.211	1.000							
(7)	Δ Cost Efficiency (%)	-0.033	0.131	0.049	0.004	-0.178	0.082	1.000						
(8)	Δ Market Sensitivity (%)	0.058	0.177	-0.081	0.193	-0.043	-0.099	-0.020	1.000					
(9)	Bank Volatility (%)	0.317	0.161	0.054	0.055	-0.012	-0.035	0.000	0.015	1.000				
(10)	Δ Index Return (%)	-0.482	-0.180	-0.062	-0.054	0.106	0.025	-0.083	0.005	-0.229	1.000			
(11)	Index Volatility (%)	0.362	0.069	0.060	0.036	-0.100	0.016	0.004	0.098	0.353	-0.294	1.000		
(12)	Yield Curve Slope	0.062	-0.099	0.054	-0.327	0.006	0.067	-0.008	-0.350	0.028	-0.079	-0.070	1.000	
(13)	Δ Sovereign Credit Rating	0.039	-0.075	0.092	-0.065	0.005	-0.002	0.022	-0.014	0.027	-0.067	-0.040	0.192	1.000

Table 5: Baseline Regression with Biodiversity

This table displays the regression results of the relationship between biodiversity and CDS change with different fixed effects and clustered standard errors. Column 1 does not include fixed effects, while column 2 controls for time fixed effects, column 3 for time and bank fixed effects and column 4 for time and country fixed effects. CDS change is depicted in % change per quarter and the biodiversity measure is the standardized quarterly average of the biodiversity index. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)			
	(1)	(2)	(3)	(4)
Intercept	1.779***	9.137***	9.420***	8.941***
	(0.244)	(1.513)	(1.442)	(1.507)
Biodiversity	1.946***	4.231***	4.231***	4.231***
	(0.507)	(0.795)	(0.808)	(0.801)
Time Fixed Effects	No	Yes	Yes	Yes
Bank Fixed Effects	No	No	Yes	No
Country Fixed Effects	No	No	No	Yes
Clustered Standard Errors	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248
Adjusted R^2	0.006	0.319	0.302	0.313

ever, the coefficients of ΔLEV, ΔAQ and ΔCE are now negative, contradicting the expectation established in Chapter 5.1. While the insignificance does not indicate any relationship in either direction, the time lag does not seem to result in a better model fit. This claim is supported by the decrease in adjusted R^2 from 0.319 to 0.317 for the regression with all balance sheet variables and time fixed effects, compared to Table 6.

Next, the remaining control variables are included in the model. The regression formula changes as follows:

$$\begin{aligned}\Delta CDS_{i,t} = \alpha &+ \beta_1 Biodiversity_t + \beta_2 \text{LEV}_{i,t} + \beta_3 \text{AQ}_{i,t} \\ &+ \beta_4 \text{Size}_{i,t} + \beta_5 \text{FS}_{i,t} + \beta_6 \text{CE}_{i,t} \\ &+ \beta_7 \text{MS}_{i,t} + \beta_8 \text{BVOL}_{i,t} + \beta_9 \text{IDXRET}_{i,t} \\ &+ \beta_{10} \text{IDXVOL}_{i,t} + \beta_{11} \text{Yield}_{i,t} + \beta_{12} \text{CR}_{i,t} \\ &+ \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t}\end{aligned} \tag{5}$$

$\Delta\text{BVOL}_{i,t}$ is the change in bank volatility, $\Delta\text{IDXRET}_{i,t}$ the change in quarterly stock market index return, $\text{IDXVOL}_{i,t}$ the stock market index volatility, $\text{Yield}_{i,t}$ the yield curve slope and $\Delta\text{CR}_{i,t}$ the change in sovereign credit rating. Again, the regression is run both with and without a time lag for the balance sheet variables, to account for a potential delayed effect. The regression formula changes accordingly and compared to formula 5, the balance sheet variables are included from

Table 6: Baseline Regression with Biodiversity and Balance Sheet Variables

This table displays the regression results of the relationship between biodiversity and CDS change including balance sheet control variables without time lag. Columns 1 to 6 show the step-by-step introduction of balance sheet variables. Columns 7 to 9 show the full regression with time fixed effects, time and bank fixed effects and time and country fixed effects, respectively. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	9.138***	9.395***	8.875***	9.097***	9.450***	9.249***	9.453***	10.154***	9.353***
	(1.515)	(1.557)	(1.495)	(1.501)	(1.616)	(1.541)	(1.698)	(1.727)	(1.691)
Biodiversity	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***
	(0.796)	(0.796)	(0.796)	(0.796)	(0.796)	(0.796)	(0.797)	(0.810)	(0.803)
ΔLEV	-0.004						0.166	0.118	0.113
	(0.830)						(0.964)	(1.012)	(0.998)
ΔAQ		0.053					0.046	0.050	0.058
		(0.045)					(0.043)	(0.047)	(0.046)
ΔSize			-4.662				-3.811	-2.636	-2.306
			(2.910)				(2.966)	(3.649)	(3.489)
ΔFS				0.029			0.039	0.028	0.027
				(0.092)			(0.097)	(0.098)	(0.095)
ΔCE					0.041		0.037	0.041	0.041
					(0.035)		(0.037)	(0.038)	(0.037)
ΔMS						0.022	0.020	0.020	0.019
						(0.016)	(0.016)	(0.018)	(0.017)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	No	No	No	Yes	No
Country FE	No	No	No	No	No	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.318	0.319	0.320	0.318	0.319	0.320	0.319	0.302	0.313

the period $t-1$. Table 8 and Table 9 show the stepwise introduction of the market variables for both the regressions with and without lagged balance sheet variables. In Table 8, the variables $\Delta BVOL$, $\Delta IDXRET$ and *IDXVOL* are all significant at the 1% significance level once introduced. When comparing to the results in column seven of Table 6, the adjusted R^2 increases from 0.319 to 0.351, 0.388, and 0.340, respectively. Additionally, ΔCR is significant at the 10% significance level and slightly increases the model fit to 0.320. All coefficients are in line with the expected sign, except *Yield* and ΔCR. With time lag, the results stay similar, as shown in Table 9. The significance levels are all equal, except for ΔMS in column three. The coefficients and standard errors of all significant variables are comparable to the results without time lag. In terms of model fit, the regression with time lag underperforms in comparison, as the addition of market variables consistently leads to a slightly lower adjusted R^2, as already shown in Table 6 and Table 7. The full regression results are presented in Table 10. Panel A and Panel B include the results without and with time lag, respectively. In Panel A, *Biodiversity* is significant at the 5% significance level across all regressions. The coefficient is positive and between 1.532 and 1.559 with a standard error of 0.702 and 0.710. This also suggests economic significance, as an increase of one standard deviation in the biodiversity index is related to an increase in CDS spread changes by 1.532 to 1.559 percentage points.

Furthermore, $\Delta BVOL$ is significant at the 1% level with a coefficient between 0.094 to 0.101 and a standard error of 0.022. The effect of $\Delta BVOL$ on CDS spreads is therefore positive and an increase of 100bp is associated with an increase in CDS spread changes of 0.094 and 0.101 percentage points. Additionally, $\Delta IDXRET$ is also significant at the 1% level across all regressions. It is negatively related to CDS spread changes, as expected. The coefficient lies around -0.896 with a standard error of around 0.093. An increase in stock index returns by 100bp is therefore related to a decrease in CDS price changes of 0.896 percentage points. Lastly, *IDXVOL* is also significant at the 1% level and positively related to CDS price changes. For the results with time fixed effects only, the coefficient is 6.042 with a standard error of 1.524. Introducing bank or country fixed effects increases both statistics. The coefficient increases to 8.593 or 8.618 and the standard error to 2.250 or 2.224. Therefore, the economic impact of an increase in stock index volatility by 100bp is an increase in CDS spread changes between 6.042 and 8.618 percentage points. In contrast, none of the balance sheet variables are significant. Additionally, the coefficients of *Yield* and ΔCR are

Table 7: Baseline Regression with Biodiversity and lagged Balance Sheet Variables

This table displays the regression results of the relationship between biodiversity and CDS change including balance sheet control variables with time lag. Columns 1 to 6 show the step-by-step introduction of balance sheet variables. Columns 7 to 9 show the full regression with time fixed effects, time and bank fixed effects and time and country fixed effects respectively. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	9.138***	9.082***	8.984***	9.152***	9.209***	9.174***	8.894***	9.676***	8.840***
	(1.517)	(1.506)	(1.601)	(1.508)	(1.564)	(1.508)	(1.643)	(1.464)	(1.636)
Biodiversity	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***	4.231***
	(0.796)	(0.796)	(0.796)	(0.796)	(0.796)	(0.796)	(0.797)	(0.810)	(0.803)
ΔLEV	0.006						-0.049	-0.109	-0.130
	(0.767)						(0.830)	(0.860)	(0.842)
ΔAQ		-0.014					-0.028	-0.031	-0.021
		(0.047)					(0.047)	(0.054)	(0.051)
ΔSize			-0.939				-1.533	-0.199	0.069
			(2.709)				(2.746)	(3.035)	(2.887)
ΔFS				-0.113			-0.126	-0.149	-0.149
				(0.096)			(0.104)	(0.114)	(0.111)
ΔCE					-0.008		-0.009	-0.006	-0.005
					(0.024)		(0.026)	(0.028)	(0.027)
ΔMS						0.006	0.005	0.004	0.004
						(0.014)	(0.014)	(0.015)	(0.015)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	No	No	No	Yes	No
Country FE	No	No	No	No	No	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.318	0.318	0.318	0.319	0.318	0.319	0.317	0.299	0.310

insignificant as well and indicate no significant relationship with CDS spread changes. The adjusted R^2 is 0.418, 0.406, and 0.415 for the regressions with the different fixed effects. Therefore, the model predicts 40.6% to 41.8% of the variance in CDS spread changes, which is comparable to Corò et al. (2013), who also study determinants of CDS spread changes. Panel B shows the results with lagged balance sheet variables. Compared to Panel A, the coefficients and standard errors of the significant variables are equivalent with only small deviations. Adding the time lag does not change the insignificance of the balance sheet variables. Concluding, there is therefore no evidence of a significant impact on CDS price changes for these variables. This result contrasts the studies of Drago et al. (2017) and Hasan et al. (2016). However, these studies employ a different empirical design, which might cause these result disparities. Both include CDS spreads in levels instead of period changes. While the balance sheet variables are shown to significantly determine the magnitude of CDS spreads in levels, this thesis shows that spread movements are not significantly affected by these measures. While Corò et al. (2013) show that changes in leverage do impact CDS price changes, they study corporate CDS in different industries. This thesis focuses on banks, which have a different liability structure where the leverage ratios are considerably higher and similar across different banks, which might explain the different results (Hasan et al., 2016). Therefore, the conclusions of this analysis do not contradict these

findings but rather add a new dimension to the analysis of bank's CDS determinants. Furthermore, these results also show that CDS spread changes are strongly influenced by the stock markets, which is in line with the findings of Norden and Weber (2009), who discover co-movements of stock and CDS markets, where stock returns lead changes in CDS spreads.

Concluding, the results of the regressions confirm *Hypothesis 1* and show that negative biodiversity news lead to increasing CDS spreads. The introduction of a time lag does not change the implications of the baseline regression. Therefore, it does not seem to capture anyother effects compared to the model without lag. Going forward, the sample without time lag will therefore serve as the basis in any further analysis.

Concluding, the results of the regressions confirm *Hypothesis 1* and show that negative biodiversity news lead to increasing CDS spreads. The introduction of a time lag does not change the implications of the baseline regression. Therefore, it does not seem to capture anyother effects compared to the model without lag. Going forward, the sample

Table 8: Stepwise Regression with all Control Variables without Time Lag

This table displays the regression results of the relationship between biodiversity and CDS change with the stepwise introduction of market control variables without time lag. Columns 1 to 5 show the step-by-step introduction of market variables with time fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)				
	(1)	(2)	(3)	(4)	(5)
Intercept	9.341***	11.021***	-1.937	8.649***	9.595***
	(1.653)	(1.824)	(2.331)	(1.556)	(1.693)
Biodiversity	2.868***	2.584***	4.077***	4.175***	4.301***
	(0.767)	(0.751)	(0.731)	(0.779)	(0.781)
ΔLEV	-0.088	0.050	0.125	0.162	0.083
	(0.913)	(0.767)	(0.867)	(0.969)	(0.929)
ΔAQ	0.038	0.022	0.039	0.058	0.048
	(0.043)	(0.043)	(0.042)	(0.043)	(0.043)
ΔSize	-4.323	-3.222	-3.400	-3.836	-3.883
	(2.729)	(2.342)	(2.723)	(2.983)	(3.000)
ΔFS	0.050	0.063	0.013	0.042	0.038
	(0.090)	(0.100)	(0.094)	(0.096)	(0.097)
ΔCE	0.040	0.001	0.026	0.036	0.036
	(0.034)	(0.036)	(0.037)	(0.037)	(0.037)
ΔMS	0.019	0.019	0.025	0.021	0.019
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
ΔBVOL	0.131***				
	(0.023)				
ΔIDXRET		-0.971***			
		(0.110)			
IDXVOL			11.562***		
			(2.501)		
Yield				0.917	
				(0.759)	
ΔCR					3.900*
					(2.264)
Time FE	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	No
Country FE	No	No	No	No	No
Clustered SE	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.351	0.388	0.340	0.319	0.320

without time lag will therefore serve as the basis in any further analysis.

6.2 Impact of the Kunming Declaration

As introduced in earlier chapters, the Kunming Declaration represents a salient event for promoting biodiversity risk awareness in recent years. It has been adopted at the 15th Conference of the Parties to the CBD (COP 15) on 13th of October in 2021 (CBD Secretariat, 2021b). In the declaration, more than 100 countries announced their commitment to increase global biodiversity conservation efforts across all economic and societal sectors (Garel et al., 2024; Kalhoro & Kyaw, 2024). Especially the finance sector was mentioned, as sustainable financial flows and their voluntary engagement are needed to implement sustainable developments (CBD Secretariat, 2021b).

Due to the unexpected severity of outcomes, COP 15 can be viewed as a shock, which increased transition risks and uncertainty for global industries (Garel et al., 2024). Using this external shock, this thesis aims to determine whether the effect of biodiversity news on CDS prices changes with the Kunming Declaration. Recent studies have similarly used the declaration to examine how it influences biodiversity risk effects (Garel et al., 2024; Kalhoro & Kyaw, 2024). Due to the potential policy changes and increased risk uncertainty, the impact of the Kunming Declaration is expected to be positive. The effect of biodiversity news on CDS prices should therefore be higher for the periods following the declaration. To

Table 9: Stepwise Regression with all Control Variables with Time Lag

This table displays the regression results of the relationship between biodiversity and CDS change with the stepwise introduction of market control variables with time lag. Columns 1 to 5 show the step-by-step introduction of market variables with time fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)				
	(1)	(2)	(3)	(4)	(5)
Intercept	8.932***	10.438***	-2.496	8.457***	8.975***
	(1.672)	(1.794)	(2.215)	(1.704)	(1.629)
Biodiversity	2.872***	2.567***	4.072***	4.201***	4.300***
	(0.767)	(0.758)	(0.730)	(0.778)	(0.781)
ΔLEV	0.135	-0.110	0.209	-0.025	-0.002
	(0.825)	(0.777)	(0.851)	(0.835)	(0.809)
ΔAQ	-0.021	-0.042	-0.044	-0.023	-0.023
	(0.048)	(0.048)	(0.045)	(0.047)	(0.049)
ΔSize	-1.317	-3.342	-0.069	-1.576	-1.973
	(2.570)	(2.562)	(2.932)	(2.773)	(2.732)
ΔFS	-0.137	-0.124	-0.132	-0.126	-0.122
	(0.099)	(0.111)	(0.104)	(0.105)	(0.106)
ΔCE	-0.012	-0.001	-0.024	-0.009	-0.009
	(0.027)	(0.024)	(0.027)	(0.026)	(0.026)
ΔMS	0.005	0.004	0.013	0.006	0.005
	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
ΔBVOL	0.131***				
	(0.024)				
ΔIDXRET		-0.981***			
		(0.107)			
IDXVOL			11.984***		
			(2.434)		
Yield				0.494	
				(0.707)	
ΔCR					3.865*
					(2.242)
Time FE	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	No
Country FE	No	No	No	No	No
Clustered SE	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.348	0.388	0.338	0.316	0.317

incorporate this event in the sample of this thesis, a binary dummy variable $Kunming_t$ is constructed, with a value of one for the period of the Kunming Declaration and onwards and zero before. Interacting this variable with the biodiversity index will show if the baseline results change due to the external shock. The regression model is the following:

$$\begin{aligned} CDS_{i,t} =& \alpha + \beta_1 Biodiversity_t + \beta_2 Kunming_t \\ &+ \beta_3 Biodiversity_t \times Kunming_t \\ &+ \beta_4 X_{i,t} + \gamma_t + \mu_i + \varphi_j + \epsilon_{i,t} \end{aligned} \tag{6}$$

where the main coefficient of interest is β_3 of the interaction term. The term $X_{i,t}$ contains all control variables of the baseline regression. The results are reported in Panel A of Table 11. After adding the interaction term, *Biodiversity* is no longer significant. This suggests that there is no evidence for a significant relationship between biodiversity news and CDS spread changes before the Kunming Declaration. However, the interaction $Biodiversity \times Kunming$ is significant at the 5% significance level, with a positive coefficient of 2.179 and a standard error of 1.068 for the regression with time fixed effects only. Adding bank or country fixed effects decreases the coefficient slightly to 2.033 and 2.034, respectively. The relationship between biodiversity news and CDS prices is therefore sufficiently higher since the Kunming Declaration. On average, banks experienced CDS price changes which were 2.033 to 2.179 percentage points higher for every standard deviation increase of the biodiversity index, after the declara-

Table 10: Baseline Regression with all Control Variables

This table displays the regression results of the relationship between biodiversity and CDS change including all control variables without and with time lag. Panel A shows the results without and panel B with time lag. Columns 1 and 4 include time fixed effects, 2 and 5 time and bank fixed effects and 3 and 6 time and country fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)					
	Panel A			Panel B		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	4.117*	3.759	2.966	4.018	4.004	3.084
	(2.315)	(2.659)	(2.595)	(2.411)	(2.550)	(2.686)
Biodiversity	1.532**	1.556**	1.559**	1.541**	1.561**	1.563**
	(0.702)	(0.710)	(0.704)	(0.704)	(0.709)	(0.702)
ΔLEV	-0.163	-0.469	-0.474	0.198	0.027	-0.001
	(0.736)	(0.735)	(0.728)	(0.775)	(0.804)	(0.786)
ΔAQ	0.025	0.023	0.033	-0.037	-0.042	-0.030
	(0.043)	(0.046)	(0.045)	(0.049)	(0.054)	(0.051)
ΔSize	-3.471	-2.851	-2.451	-2.346	-1.365	-1.019
	(2.156)	(2.609)	(2.497)	(2.454)	(2.601)	(2.487)
ΔFS	0.058	0.035	0.033	-0.135	-0.162	-0.163
	(0.093)	(0.097)	(0.094)	(0.106)	(0.117)	(0.113)
ΔCE	-0.000	0.006	0.006	-0.013	-0.007	-0.006
	(0.034)	(0.036)	(0.035)	(0.024)	(0.025)	(0.025)
ΔMS	0.023	0.017	0.016	0.010	0.006	0.005
	(0.017)	(0.020)	(0.020)	(0.015)	(0.017)	(0.016)
ΔBVOL	0.101***	0.095***	0.094***	0.101***	0.095***	0.095***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
ΔIDXRET	-0.896***	-0.896***	-0.895***	-0.903***	-0.907***	-0.906***
	(0.093)	(0.093)	(0.092)	(0.092)	(0.094)	(0.093)
IDXVOL	6.042***	8.593***	8.618***	6.109***	8.420***	8.437***
	(1.524)	(2.250)	(2.224)	(1.564)	(2.317)	(2.290)
Yield	0.853	1.065	1.085	0.589	0.716	0.727
	(0.628)	(1.014)	(0.997)	(0.560)	(0.831)	(0.822)
ΔCR	0.064	0.331	0.325	0.075	0.248	0.232
	(2.609)	(2.696)	(2.671)	(2.452)	(2.543)	(2.522)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	Yes	No	No	Yes	No
Country FE	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.418	0.406	0.415	0.417	0.406	0.415

tion. The lack of significance of the *Biodiversity* variable may either be caused by a shortcoming in the model, which may not capture the existing effect, or correctly indicate its non-existence in the pre-Kunming period. The latter would be in line with the findings of Garel et al. (2024), who employ a different biodiversity measure and find increasing effects on stock returns only in the post-Kunming period. Furthermore, Coqueret et al. (2025) show similar results for the effect of biodiversity risks on asset prices. While not specifically using the Kunming Declaration as an external shock, they find that a negative premium has emerged since the year 2021. Concluding, the findings provide evidence in favour of the confirmation of *Hypothesis 2*. Following similar research, the results further indicate that the significant effect of biodiversity news on CDS spread changes has only occurred since the Kunming Declaration in 2021. Therefore, the relation proposed in *Hypothesis 1* and confirmed in Chapter 6.1 does not seem to hold for the full sample and is only evident in the post-Kunming period.[4]

Adding time fixed effects to a model with a time-dependent binary variable seems problematic at first glance, as they would normally absorb any invariant variable within a time period. However, since the time fixed effects are employed on a higher frequency, any variation within a half-year period is not absorbed. The binary *Kunming* variable switches to 1 in Q4 2021, as the declaration was signed in early October and therefore, there exists variation within a half-year period. Consequently, the variables, as well as the interaction, are not absorbed. To provide further transparency on this approach, Panel B reports the regression results without time fixed effects. Importantly, the interaction term is not affected and is comparable to the results obtained in Panel A. Furthermore, columns four and five show the importance of time fixed effects, as time shocks and trends are captured by the biodiversity and non-reported control variables in their absence. Using the half-yearly time fixed effects therefore improves the model's accuracy, as shown by the higher adjusted R^2 of the regressions in Panel A.

In the subsequent chapters, further country and bank characteristic will be included as interactions to test for heterogeneous cross-sectional effects. The regression will hereby differ between the two-way interactions with the biodiversity variable and a three-way interaction with the Kunming variable in addition. While the regression results in Table 11 suggest that the significant impact of biodiversity news on CDS prices only occurs after the Kunming Declaration, there might be mediating effects, which result in significant relationships prior to it. Therefore, the two-way interactions will be implemented to capture such effects over the full sample period. However, given the results from Table 11, it might also be the case that the relation is only occurring after the Kunming Declaration. In this case, bank and country characteristics might still instigate heterogeneous results across sub-groups in the sample, around the declaration. These will be captured by the three-way interactions. However, the inclusion of these regressions is not without problems. When interacting two variables X_1 and X_2, the interaction term $X_1 \times X_2$ is naturally somewhat collinear with the variables it is constituted of (Tate, 1984). This dependency between two or more variables is called multicollinearity, which given its presence violates OLS assumptions and obstructs its application (Alin, 2010; Wooldridge, 2020). The problems created by multicollinearity most importantly affect standard errors, which increase in magnitude and are more sensitive to small changes in the model (K. W. Smith & Sasaki, 1979). While the overall model significance is less affected, hypothesis testing for regression coefficients suffers from inflated p-values and possibly insignificant results in the presence of multicollinearity (Alin, 2010). Approaches to counteract these problems are often hard to realise, according to Wooldridge (2020). Increasing the sample size reduces uncertainty in standard errors but is constrained by data availability. Furthermore, excluding variables from the model may lead to biased estimates. A common approach to measure and become aware of potential issues is the Variance Inflation Factor (VIF), where a larger value of VIF_i indicates some form of involvement of variable i in linear dependencies (Alin, 2010). Commonly, a VIF greater than 10 is considered as problematic, but does not immediately invalidate results, as the cutoff value is determined rather arbitrarily (Alin, 2010; Wooldridge, 2020). For the three-way interactions in the following chapters, high VIF values are detected for some of the regressions and reported in Appendix A.6. The results are still included but should therefore be regarded with caution.

6.3 A Country's State of Biodiversity

As the prior regression indicates the existence of a significant relationship between biodiversity and changes in CDS prices since the Kunming Declaration, there might also exist dependencies with country characteristics, which change this relationship. Giglio et al. (2024) show that for sovereign CDS, a country's biodiversity characteristics and presence of natural capital change the magnitude of reactions to biodiversity news in CDS pricing. Similarly, this relation might exist on a bank level, as proposed in *Hypothesis 3*. To test this effect, the state of biodiversity within a country will be introduced to the model. As a measure for the state of biodiversity, the Environmental Performance Index (EPI) is implemented. Introduced by the Yale Center for Environmental Law & Policy, it captures the state of sustainability for countries around the world. It spans 58 performance indicators across 11 environmental pillars for more than 180 countries (Block et al.,

[4] The convention in Kunming was the first part of COP 15, with the second part taking place in December 2022 in Montreal. As a result, the Kunming-Montreal Global Biodiversity Framework was signed, where the intentions of the Kunming Declaration were formulated as specific targets. Garel et al. (2024) show that using the Montreal Agreement, instead of the Kunming Declaration, as an external shock, does not have the same effect on stock returns. They argue that the outcomes were possibly more widely expected and hence did not provide new information on firms' transition risks. While not reported, this thesis similarly finds that the Montreal Agreement did not affect CDS prices, contrary to the Kunming Declaration. However, it is necessary to note that the sample only includes relatively few observations after the Montreal Agreement was signed and may therefore lack explanatory power.

Table 11: Effect of Biodiversity News on CDS Price Changes following the Kunming Declaration

This table displays the regression results of the relationship between biodiversity and CDS change including the interaction of the Kunming dummy variable. Columns 1 to 3 show the full regression with time fixed effects, time and bank fixed effects and time and country fixed effects respectively. The dependent variable is the quarterly change in CDS prices. All control variables from the baseline regression are included. Clustered standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)				
	Panel A			Panel B	
	(1)	(2)	(3)	(4)	(5)
Intercept	3.871	3.666	2.873	-7.827***	-8.938***
	(2.401)	(2.806)	(2.737)	(1.856)	(1.676)
Biodiversity	0.176	0.286	0.288	-1.034*	-1.031*
	(1.152)	(1.173)	(1.163)	(0.569)	(0.564)
Kunming	4.388	4.119	4.115	4.587***	4.631***
	(3.314)	(3.405)	(3.373)	(1.020)	(1.009)
Biodiversity × Kunming	2.179**	2.033*	2.034*	2.156**	2.162**
	(1.068)	(1.072)	(1.062)	(0.849)	(0.840)
Control Variables	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	No	No
Bank FE	No	Yes	No	Yes	No
Country FE	No	No	Yes	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.418	0.407	0.416	0.321	0.331

2024). Each indicator receives a score and weight, which when aggregated form the overall EPI score. The EPI covers three main scopes, which are climate change, environmental health, and ecosystem vitality. For the analysis, the score of ecosystem vitality is used, which consists of indicators spanning the topics biodiversity & habitat, forests, fisheries, air pollution, agriculture, and wastewater. Through various indicators, it measures how well a country is managing their natural resources, while conserving biodiversity and ecosystems (Block et al., 2024). A higher score indicates a better ecosystem vitality for a given country. The EPI scores are available over different time periods. However, the latest publication from the year 2024 is used for the analysis in this chapter, as it contains new metrics related to the goals of the Kunming-Montreal global biodiversity framework (Block et al., 2024). The sign of the interaction term is expected to be positive, as negative news about biodiversity should have more dominant effects on CDS spreads in countries where ecosystem degradation is more advanced.[5]

The summary statistics of the global and unfiltered data are depicted in Table 12. The state of biodiversity variable is available for 181 countries and can range between zero and 100, where higher scores indicate better states of biodiversity. The mean is at 51.12 and the median at 49.60 with a standard deviation of 13.08. Furthermore, the values span between 22.70 and 83.10, showing that there is neither a country with a flawless state of biodiversity nor a country with a fully depleted one. In the regression, the state of biodiversity is included as a binary variable. This allows for a distinction between countries with superior and inferior biodiversity conditions, compared on a global scale. It is split at the median, where the variable takes on the value of one for countries with the inferior state of biodiversity. The binary variables are determined based on the entire available data, to avoid some countries being incorrectly allocated due to the limited number of countries in the sample. Determining the median based on all available country data avoids the possibility that a country with favorable biodiversity conditions is treated contrary in the regression and vice versa. After determining the binary variable, it is matched with the rest of the sample based on the location of a bank's headquarters. In the sample, seven banks are located in countries with a score below the median of 49.60. Details are provided in Appendix A.7.

The aim of the regression is to test the interaction effect of the introduced country characteristics and biodiversity on changes in CDS prices around the Kunming Declaration. Therefore, the three-way interaction between *Biodiversity*, *Kunming* and the country variable is estimated. The econometric model is as follows:

[5] Giglio et al. (2024) also employ a country's natural capital dependency to test for heterogeneous effects between biodiversity risk and CDS prices. For this thesis, the number of countries which have a high dependency is very low, as only three of 39 banks have scores above the median. This results in few observations and high potential for multicollinearity, which limits the robustness of results. Therefore, this measure is not included in the scope of this thesis.

Table 12: Summary Statistics of the State of Biodiversity Measure

This table shows the summary statistics for the state of biodiversity measure.

Summary Statistics								
Country Factors	N	Mean	St. Dev.	Min	25th perc.	Median	75th perc.	Max
State of Biodiversity	181	51.12	13.08	22.70	41.80	49.60	60.40	83.10

$$\begin{aligned} CDS_{i,t} =& \alpha + \beta_1 Biodiversity_t + \beta_2 Kunming_t \\ &+ \beta_3 BioState_{med,i} + \beta_4 Biodiversity_t \\ &\times Kunming_t + \beta_5 Biodiversity_t \\ &\times BioState_{med,i} + \beta_6 Kunming_t \\ &\times BioState_{med,i} + \beta_7 Biodiversity_t \\ &\times Kunming_t \times BioState_{med,i} \\ &+ \beta_8 X_{i,t} + \gamma_t + \epsilon_{i,t} \end{aligned} \tag{7}$$

where $BioState_{med,i}$ is the state of biodiversity variable. To provide robustness, the effect of the country characteristic is also estimated separately over the full sample, to show its effect independent of the Kunming Declaration. In terms of fixed effects, only time fixed effects are included, as bank or country fixed effects would fully absorb the time-invariant country characteristic. Table 13 reports the result of the regression. Column one shows the statistics of the regression with the two-way interaction and column two of the regression with the triple interaction. In column one, the coefficient of *Biodiversity* is positive and significant at the 5% significance level. The magnitude of the coefficient and standard error are slightly higher compared to the baseline regression presented in Table 10. $BioState_{med}$ is significant at the 5% significance level as well, indicating a relationship independent of *Biodiversity*, where CDS price changes were lower for countries with a worse state of biodiversity on average. The interaction term is insignificant and the standard error is rather high, which suggests substantial uncertainty and does not support any confident conclusion about an effect in either direction. With the introduction of *Kunming* in column two, *Biodiversity* becomes insignificant. Compared to Table 11, the interaction $Biodiversity \times Kunming$ is now insignificant, and with a p-value of 0.133 above the 10% significance threshold. Therefore, there is no resilient evidence for the existence of an effect of the interaction $Biodiversity \times Kunming$. Neither the interaction $Biodiversity \times BioState_{med}$, nor the triple interaction are significant. Hence, there is no indication that the effect of the Kunming Declaration is dependent on a country's state of biodiversity. *Hypothesis 3* can therefore not be confirmed, given the analysis for the sample in this thesis. The model fails to capture previously observed significant effects, which may be caused by two problems. First, the model may lack predictive power after introducing the triple interaction, since the sample is essentially split into four subgroups.[6] The sample size may therefore be too small to estimate a significant heterogeneous effect. Furthermore, the VIFs are exceeding the threshold of 10 in some cases, which may affect the results, as discussed in Chapter 6.2. Overall, the results show no evidence that CDS prices react differently to biodiversity news in countries with higher biodiversity degradation, neither before nor after the Kunming Declaration. Importantly, the results for $Biodiversity \times Kunming$ and $Biodiversity \times Kunming \times BioState_{med}$ do not necessarily contradict the effects obtained in Table 11. The lack of significance only fails to provide evidence for the according relations in this particular regression but does not prove their non-existence. To show robustness, the ecosystem vulnerability score from the Notre Dame Global Adaptation Initiative (ND-GAIN) database is used to provide an alternative measure of a country's ecosystem health. Regression results are disclosed in Chapter 7 and confirm the findings.

6.4 Membership in Bank Initiatives

As established in Chapter 2.2, the role of banks in the context of biodiversity loss has been increasingly acknowledged in recent years. In response to the emerging challenges, several banking initiatives have been formed, which aim to call for action and support collaboration among institutions. For example, the Finance for Biodiversity (FfB) Foundation has introduced a pledge, where all signatories aim to take on responsibility and support the conservation of biodiversity (FfB, 2024). In September 2023, the Task Force on Nature-related Financial Disclosures (TNFD) published a recommendation framework, which supports decision-making processes, risk management, and disclosure of nature-related risks. They urge the adoption of this framework to address insufficient risk management practices facing accelerating nature loss (TNFD, 2023). Furthermore, the UN Environment Program Finance Initiative (UNEP FI) has launched the Principles of Responsible Banking in 2019. While not exclusively focusing on biodiversity, the framework encourages sustainable finance in the areas of nature, climate, and circular economy among others (UNEP FI, 2024). More importantly, the

[6] Both binary variables *Kunming* and $BioState_{med}$ take on either zero or one. The triple interaction therefore estimates the four cases (0,0), (0,1), (1,0), and (1,1), where each bank is included in a subgroup based on its characteristics. Results are the coefficients of $Biodiversity$, $Biodiversity \times Kunming$, $Biodiversity \times BioState_{med}$ and $Biodiversity \times Kunming \times BioState_{med}$, respectively.

Table 13: Cross-Sectional Effect of Countries' State of Biodiversity

This table displays the regression results of the relationship between biodiversity and CDS change depending on the state of biodiversity and the Kunming Declaration. CDS change is depicted in % change per quarter and the biodiversity measure is the standardized quarterly average of the biodiversity index. The binary variable BioState_{med} represents the state of biodiversity. Panel A includes the interaction effect of the state of biodiversity for the full sample and panel B dependent on the Kunming Declaration. All regressions include control variables and half-yearly time fixed effects. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)	
	(1)	(2)
Intercept	4.607*	4.172*
	(2.277)	(2.454)
Biodiversity	1.688**	0.473
	(0.660)	(1.172)
Kunming		4.454
		(3.263)
Biodiversity × Kunming		1.863
		(1.212)
BioState_{med}	-2.399**	-2.195*
	(1.003)	(1.168)
Biodiversity × BioState_{med}	-0.626	-1.135
	(1.140)	(0.986)
Kunming × BioState_{med}		-0.817
		(1.548)
Biodiversity × Kunming × BioState_{med}		1.386
		(2.035)
Control Variables	Yes	Yes
Time FE	Yes	Yes
Bank FE	No	No
Country FE	No	No
Clustered SE	Yes	Yes
N	1,248	1,248
Adj. R^2	0.418	0.418

framework provides guidance to promote setting biodiversity targets, to push their inclusion in banks' overall strategies, and to allocate responsibilities (UNEP FI, 2022). Banks are encouraged to halt financing of activities which are harmful to biodiversity and conduct robust risk mitigation and due diligence. The Principles for Responsible Banking are therefore an important step towards promoting biodiversity risk awareness in the banking sector (Hudson, 2024). By advocating these initiatives, banks openly acknowledge and engage with the potential issues of biodiversity risks. In recent years, studies have shown such decisions to impact the market perception of credit risks. Focusing on climate risks, Kölbel et al. (2022) and Carbone et al. (2021) both show that CDS markets react to the disclosure of climate risks twofold. While CDS spreads might rise due to increased risk perception, they might also decrease, since the disclosure lowers uncertainty. Furthermore, Kalhoro and Kyaw (2024) find that firms who actively manage biodiversity risks experience positive effects. Therefore, banks' participation in biodiversity initiatives is used to analyze whether they experience diverging effects of biodiversity risks on CDS prices, as proposed in *Hypothesis 4*. The above-mentioned initiatives are considered, as they specifically target financial institutions. The TNFD framework is ruled out, as banks have only started to adopt its recommendations as early as 2024, which lies beyond the scope of this thesis (TNFD, 2023). Furthermore, the FfB pledge has only one overlap between signatories and banks in the sample, which leads to insufficient observations. Therefore, only the UNEP FI initiative is considered. For each bank, a dummy variable is constructed, which is one for "green banks", which have joined the programs and zero for those which have not. The variable is constructed as time series data, which only includes a bank as green from the moment it joined. The members of the initiative and their date of adoption are disclosed on the UNEP FI website.[7] Of the total 39 banks, 18 have either been members over the full time period of the sample or have joined during. Analogous to the approach taken in chapter 6.3, the overall effect, as well as the effect before and after the Kunming Declaration, are tested. The regression model is the following:

[7] https://www.unepfi.org/members/.

$$\begin{aligned} CDS_{i,t} =& \alpha + \beta_1 Biodiversity_t + \beta_2 Kunming_t \\ &+ \beta_3 Green_{i,t} + \beta_4 Biodiversity_t \times Kunming_t \\ &+ \beta_5 Biodiversity_t \times Green_{i,t} \\ &+ \beta_6 Kunming_t \times Green_{i,t} \\ &+ \beta_7 Biodiversity_t \times Kunming_t \times Green_{i,t} \\ &+ \beta_8 X_{i,t} + \gamma_t + \epsilon_{i,t} \end{aligned} \tag{8}$$

where $Green_{i,t}$ is the introduced dummy variable for the banks which are members of the UNEP FI initiative. Table 14 shows the results of the regression. In column one, the baseline regression is extended by the interaction $Biodiversity \times Green$. The coefficient of *Biodiversity* is significant at the 5% significance level, with a slightly higher coefficient of 2.078 compared to the baseline regression. This suggests a significant effect for non-green banks, where a one standard deviation increase of the biodiversity index is related to an increase in CDS price changes of 2.078 percentage points. While the interaction term is negative, which would suggest lower effects of biodiversity news on CDS prices, it is insignificant. Therefore, there is no evidence that the observed effect of *Biodiversity* on non-green banks is divergent from green banks. Additionally, the adjusted R^2 is slightly lower compared to the baseline regression, which indicates that the interaction does not add any information to the model. For the triple interaction with Kunming, the results are given in column two. The coefficient of *Biodiversity* is insignificant, in contrast to the interactions $Biodiversity \times Green$ and $Biodiversity \times Kunming \times Green$, which are significant at the 5% and 10% significance level, respectively. Before the Kunming Declaration, green banks therefore experienced lower effects of biodiversity news on their CDS spreads, indicated by the negative coefficient of $Biodiversity \times Green$. This result is as expected and shows that acknowledging and engaging in risk management helps to reduce negative impacts on credit risk, in line with the uncertainty reduction effect hypothesized by Kölbel et al. (2022). After the Kunming Declaration, this effect switches. The positive coefficient of the interaction $Biodiversity \times Kunming \times Green$ indicates that CDS spreads of green banks have been impacted more strongly by biodiversity news in the post-Kunming period. This change is likely due to a higher risk awareness for the severe urgency and impacts of biodiversity risks after the Kunming Declaration raised awareness (Kölbel et al., 2022). For non-green banks, there is no indication of significant effects, as *Biodiversity* and $Biodiversity \times Kunming$ are insignificant. In conclusion, the regression results provide evidence that the relationship between biodiversity news and CDS prices is significantly different for banks who actively acknowledge and engage with their biodiversity risk exposure. Before the Kunming Declaration, green banks experience decreasing effects on CDS prices, which is in line with the uncertainty reduction effect. Afterwards, this relationship changes and green banks experience increasing effects, in line with the risk perception effect (Kölbel et al., 2022). The insignificance for non-green banks does not necessarily lead to the conclusion of a non-existence of any relationship. Rather, the model does not provide evidence for one, which may be caused by the sample distribution. Non-green banks are in large part located in the US, which may cause biased estimates. Note, the VIFs for the triple interaction are significantly lower compared to previous regressions. Only the VIF of Kunming is surpassing the threshold of 10, all others are below. Paired with moderate standard errors, the issue of multicollinearity therefore seems less threatening for this three-way interaction, but the results should still be treated with caution. Concluding, the results do not provide evidence for the confirmation of *Hypothesis 4*, given the regression over the full sample. Analyzing the effects for the pre- and post-Kunming period separately provides the expected results, which are in need of confirmation outside this thesis.

6.5 Impact on US Banks

Motivated by *Hypothesis 5*, another cross-section of the sample is analysed. Given that the USA is not officially part of the CBD, the Kunming Declaration may have diverging effects on the US economy, compared to countries which actively contributed to it. As the declaration may encourage policies and regulations, transition risks faced by banks and firms in the USA may be sufficiently lower (Garel et al., 2024). Consequently, banks' credit risk may be less affected by biodiversity risks. To test this hypothesis, a cross-sectional approach is used, by dividing the sample into banks located in the USA and those which are not. Following previous approaches, a binary dummy variable is introduced. The variable USA_i takes on the value one for every bank which is headquartered in the USA.

The regression formula is the following:

$$\begin{aligned} CDS_{i,t} =& \alpha + \beta_1 Biodiversity_t + \beta_2 Kunming_t + \beta_3 USA_i \\ &+ \beta_4 Biodiversity_t \times Kunming_t \\ &+ \beta_5 Biodiversity_t \times USA_i + \beta_6 Kunming_t \\ &\times USA_i + \beta_7 Biodiversity_t \times Kunming_t \\ &\times USA_i + \beta_8 X_{i,t} + \gamma_t + \epsilon_{i,t} \end{aligned} \tag{9}$$

The regression includes half-yearly time fixed effects, clustered standard errors, and the full set of control variables $X_{i,t}$. Table 15 presents the results of the regression. In column one, the results without the *Kunming* variable are shown. The coefficient of *Biodiversity* is positive and significant at the 10% significance level with a value of 1.410. Its standard error is slightly higher compared to the baseline regression with 0.777, indicating higher uncertainty of the estimate. The coefficient of USA_i is positive and significant at the 5% level, with a coefficient of 1.648. This result indicates that CDS spread changes where on average higher compared to non-US banks, independent of *Biodiversity*. The coefficient of the interaction is not significant, which suggests no evidence for heterogeneous effects of biodiversity news on CDS prices for US and non-US banks. Column two shows the results with the *Kunming* interaction. The coefficient of *Biodiversity* is insignificant.

Table 14: Cross-Sectional Effect of Sustainable Initiative Membership

This table displays the regression results of the relationship between biodiversity and CDS change including the interaction of the green bank and Kunming Declaration dummy variables *Green* and *Kunming*. Column 1 includes the regression without and column 2 with the Kunming interaction. CDS change is depicted in % change per quarter and the biodiversity measure is the standardized quarterly average of the biodiversity index. Both regressions include control variables and half-yearly time fixed effects. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)	
	(1)	(2)
Intercept	4.141*	4.292*
	(2.338)	(2.421)
Biodiversity	2.078**	1.245
	(0.846)	(1.258)
Kunming		4.003
		(3.525)
Biodiversity × Kunming		0.604
		(1.461)
Green	0.444	0.094
	(0.657)	(0.751)
Biodiversity × Green	-1.087	-2.626**
	(0.911)	(1.199)
Kunming × Green		0.637
		(1.273)
Biodiversity × Kunming × Green		3.529*
		(1.897)
Control Variables	Yes	Yes
Time FE	Yes	Yes
Bank FE	No	No
Country FE	No	No
Clustered SE	Yes	Yes
N	1,248	1,248
Adj. R^2	0.417	0.418

All three interaction terms $Biodiversity \times Kunming$, $Biodiversity \times USA$, and $Biodiversity \times Kunming \times USA$ are significant at the 1% level. First, the positive coefficient of 3.485 of $Biodiversity \times Kunming$ indicates that non-US banks have experienced stronger effects of biodiversity news on CDS price changes following the Kunming Declaration. The coefficient is economically significant, as for every standard deviation increase in the biodiversity index, CDS spreads are expected to increase by an additional 3.485 percentage points. Compared to the base Kunming interaction in Table 11 the coefficient is therefore substantially higher. Furthermore, the interaction of $Biodiversity \times USA$ is also positive with a coefficient of 2.593. It suggests that banks in the USA experienced stronger impacts of biodiversity news in the pre-Kunming period. In contrast, the coefficient of the interaction $Biodiversity \times Kunming \times USA$ is negative and with a value -5.190 of high economic significance. The result provides evidence that US banks experienced substantially lower effects of biodiversity news on CDS spreads following the Kunming Declaration. This may be caused by the expectation of no or less stringent policy and regulation changes following the global biodiversity framework compared to other countries, since the USA did not commit to the results of COP 15 as a non-member of the CBD. This result implies that transition risk plays a substantial part in the pricing of biodiversity risk in CDS spreads. A similar effect has been found by Kölbel et al. (2022), who study the effects of climate-related transition risks around the Paris Agreement. While CDS of firms in the USA were initially increasing, the withdrawal from the agreement by Donald Trump had decreasing effects on CDS, as the prospects of transition risk and new regulation disappeared. Concluding, the presented results lead to a confirmation of *Hypothesis 5*, as US banks face lower impacts of biodiversity risks after the Kunming Declaration, compared to non-US banks. Additionally, the results lead to a change in the confirmation of *Hypothesis 2*. While the Kunming Declaration did increase the impact of biodiversity news on CDS spreads, this effect does not persist for all cross-sections in the sample as banks located in the USA experienced contrary effects. *Hypothesis 2* can consequently only be confirmed for banks outside the USA.

To provide transparency, the VIFs for the three-way interaction are at times above the threshold of 10. While this might not have altered the observed effects, they should be

Table 15: Cross-Sectional Effect for US and non-US Banks

This table displays the regression results of the relationship between biodiversity and CDS change including the interaction of the USA and Kunming Declaration dummy variables *USA* and *Kunming*. Column 1 includes the regression without and column 2 with the Kunming interaction. CDS change is depicted in % change per quarter and the biodiversity measure is the standardized quarterly average of the biodiversity index. Both regressions include control variables and half-yearly time fixed effects. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)	
	(1)	(2)
Intercept	2.925	3.015
	(2.333)	(2.435)
Biodiversity	1.410*	-0.495
	(0.777)	(1.201)
Kunming		5.027
		(0.370)
Biodiversity × Kunming		3.485***
		(1.232)
USA	1.648**	2.261***
	(0.793)	(0.605)
Biodiversity × USA	0.508	2.593***
	(1.005)	(0.947)
Kunming × USA		-2.240
		(1.723)
Biodiversity × Kunming × USA		-5.190***
		(1.820)
Control Variables	Yes	Yes
Time FE	Yes	Yes
Bank FE	No	No
Country FE	No	No
Clustered SE	Yes	Yes
N	1,248	1,248
Adj. R^2	0.418	0.420

treated with caution. Currently left unexplained is the effect of biodiversity news on CDS prices for US banks before the Kunming Declaration. A comparable result has not yet been obtained in similar studies, to the current knowledge of the author. The following explanation of a potential cause is therefore not empirically observed and should be treated as such. According to the National Oceanic and Atmospheric Administration (NOAA), the decade between 2010 and 2019 has been a landmark period of climate disasters (A. B. Smith, 2020). According to A. B. Smith (2020), the USA experienced twice as many billion-dollar disasters, four of the five most costly natural disasters, as well as the two most severe wildfire seasons in USA history in the 2010s. Especially, the years 2016 and 2018 were hit by severe catastrophes, which resulted in sustained damages of several billion US dollars. In the year 2018, the USA was impacted by 14 separate billion-dollar disasters, representing the 4th highest within a year, shortly after the year 2016 with 15 separate catastrophes (A. B. Smith, 2019). Matching these periods of severe environmental destruction with the biodiversity index shows that they coincide with predominantly negative biodiversity news, as shown in Figure 2. Giglio et al. (2023) illustrate that natural disasters like the California wildfires in 2018 can lead to substantial negative news coverage about biodiversity. Furthermore, Noth and Schüwer (2023) show that natural disasters negatively impact stability, credit quality, and profitability of banks in the USA. Given these dependencies, there is an argument to be made that the stronger effect of biodiversity news on credit risk for banks in the USA before the Kunming Declaration is caused by a period of unprecedented natural disasters.

7 Robustness Checks

In this chapter, several tests are included to provide robustness for the results presented in Chapter 6. First, the baseline regression is run with the absolute value of asset quality instead of its period change. In Chapter 5.2, the data of the sample was tested for stationarity to address time trends within the variables. Asset quality was transformed to the period change in the baseline sample despite being stationary already in one of the employed unit root tests. Therefore, Table 16 depicts the baseline regression with all control variables, including asset quality in levels instead of changes.

Table 16: Robustness Check with Asset Quality as Absolute Value

This table provides robustness checks for the baseline regression presented in Chapter 6 both with and without time lag. Panel A shows the results without time lag for the balance sheet variables and Panel B with time lag. Columns 1 and 4 show the regression with time fixed effects. Columns 2 and 5 show time and bank fixed effects and columns 3 and 6 time and country fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)					
	Panel A			Panel B		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	3.879	4.715*	3.069	3.690	5.301**	3.544
	(2.329)	(2.592)	(2.523)	(2.421)	(2.410)	(2.559)
Biodiversity	1.552**	1.562**	1.560**	1.559**	1.571**	1.569**
	(0.702)	(0.712)	(0.707)	(0.704)	(0.710)	(0.703)
ΔLEV	-0.238	-0.493	-0.502	0.189	0.048	0.012
	(0.717)	(0.731)	(0.720)	(0.766)	(0.777)	(0.768)
AQ	-0.371	-0.315	-0.036	-0.443*	-0.506	-0.297
	(0.287)	(0.537)	(0.471)	(0.255)	(0.393)	(0.348)
ΔSize	-3.626	-3.098	-2.701	-2.240	-1.290	-0.943
	(2.171)	(2.594)	(2.509)	(2.506)	(2.605)	(2.489)
ΔFS	0.049	0.027	0.021	-0.124	-0.150	-0.155
	(0.097)	(0.100)	(0.098)	(0.109)	(0.119)	(0.115)
ΔCE	0.000	0.006	0.006	-0.011	-0.005	-0.005
	(0.034)	(0.035)	(0.034)	(0.024)	(0.026)	(0.025)
ΔMS	0.023	0.017	0.016	0.010	0.005	0.004
	(0.018)	(0.021)	(0.020)	(0.015)	(0.017)	(0.016)
ΔBVOL	0.099***	0.095***	0.095***	0.097***	0.094***	0.094***
	(0.021)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
ΔIDXRET	-0.888***	-0.896***	-0.898***	-0.888***	-0.899***	-0.901***
	(0.091)	(0.093)	(0.092)	(0.091)	(0.095)	(0.093)
IDXVOL	6.976***	8.739***	8.561***	7.279***	8.823***	8.675***
	(1.891)	(2.349)	(2.314)	(1.852)	(2.352)	(2.322)
Yield	0.869	0.954	0.954	0.956	0.824	0.792
	(0.642)	(0.984)	(0.970)	(0.622)	(0.833)	(0.826)
ΔCR	-0.194	0.248	0.352	-0.151	0.148	0.184
	(2.628)	(2.779)	(2.745)	(2.437)	(2.579)	(2.552)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	Yes	No	No	Yes	No
Country FE	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
N	1,248	1,248	1,248	1,248	1,248	1,248
Adj. R^2	0.418	0.406	0.415	0.417	0.406	0.415

Panel A reports the regression without, and Panel B with the half-year time lag. Comparing the results to Table 10 shows no difference regarding the significant variables, as the coefficients and standard errors remain very similar. Briefly, *AQ* is significant at the 10% level in column 4, but not consistently across regressions. Additionally, the model fit stays similar across all columns as well. The potential issue of non-stationarity was therefore correctly evaluated, as changing the variable did not alter the results and ensured stationarity in the face of unclear results in the unit root test.

Next, a robustness check is provided to account for data irregularities which appeared during the construction of the dataset. The sample includes two banks, where the fiscal year is misaligned by one quarter. Therefore, their financial reports are not covering the same period compared other banks in the sample and their balance sheet variables are hence not perfectly in line with the rest of the variables. Additionally, stock data is not available in the home currency for one bank and is therefore included in US dollars instead, which may expose the data to exchange rate movements. These banks are included to provide a more comprehensive sample. To show that the data irregularities do not distort the regression results, Table 17 shows the baseline regression for the cross-section of the sample excluding the banks with diverging data availability. Again, Panel A does not include a time lag for the balance sheet variables, while Panel B does. Comparing the results to the baseline regression in Table 10 shows similar coefficients in terms of magnitude and significance. Only *Biodiversity* is significant at the 10% instead of the 5% significance level, while also having a lower coefficient compared to the baseline regression. Notably, $\Delta Size$ is briefly significant in column one, but not consistently across all regressions. Furthermore, the model fit is consistently lower compared to the baseline regression, indicating a smaller explanatory power of the model. The inclusion of the three banks has therefore positive effects for the model fit and coefficient robustness.

To provide robustness for the regression with country characteristics in Chapter 6.3, the ND-GAIN Ecosystem Vulnerability indicator is used to provide an alternative measure of a country's state of biodiversity. It assesses ecosystem vulnerability through the exposure, sensitivity, and adaptive capacity to changing nature conditions (Chen et al., 2024). It is one of six vulnerability components, which constitute the overall ND-GAIN index. The ecosystem vulnerability combines data on changes in biome distribution and marine biodiversity, dependency on natural capital, ecological footprint, share of protected biomes, and engagement in international conventions (Chen et al., 2024). Therefore, the score pro-

Table 17: Robustness Check excluding Banks with deviating Data Availability

This table provides robustness checks for the baseline regression presented in Chapter 6 both with and without time lag. Panel A shows the results without time lag for the balance sheet variables and Panel B with time lag. Columns 1 and 4 show the regression with time fixed effects. Columns 2 and 5 show time and bank fixed effects and columns 3 and 6 time and country fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered Standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)					
	Panel A			Panel B		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	5.201**	4.934*	4.149	4.998*	4.985*	4.128
	(2.443)	(2.811)	(2.719)	(2.534)	(2.675)	(2.825)
Biodiversity	1.336*	1.360*	1.364*	1.340*	1.356*	1.356*
	(0.764)	(0.772)	(0.765)	(0.765)	(0.770)	(0.763)
ΔLEV	-0.285	-0.623	-0.644	0.259	0.063	0.023
	(0.858)	(0.865)	(0.856)	(0.931)	(0.983)	(0.961)
ΔAQ	0.031	0.035	0.045	-0.028	-0.028	-0.017
	(0.049)	(0.051)	(0.050)	(0.052)	(0.057)	(0.054)
ΔSize	-4.750**	-4.216	-3.740	-2.676	-1.792	-1.399
	(2.255)	(2.787)	(2.660)	(2.715)	(2.966)	(2.829)
ΔFS	0.046	0.024	0.022	-0.115	-0.138	-0.140
	(0.098)	(0.102)	(0.099)	(0.110)	(0.121)	(0.117)
ΔCE	0.014	0.020	0.021	-0.020	-0.015	-0.014
	(0.036)	(0.038)	(0.037)	(0.025)	(0.026)	(0.026)
ΔMS	0.018	0.011	0.010	0.009	0.005	-0.004
	(0.018)	(0.020)	(0.020)	(0.015)	(0.017)	(0.017)
ΔBVOL	0.099***	0.093***	0.092***	0.098***	0.092***	0.092***
	(0.023)	(0.024)	(0.024)	(0.023)	(0.024)	(0.024)
ΔIDXRET	-0.892***	-0.892***	-0.892***	-0.904***	-0.909***	-0.908***
	(0.098)	(0.098)	(0.098)	(0.097)	(0.099)	(0.099)
IDXVOL	5.572***	8.141***	8.183***	5.755***	7.915***	7.928***
	(1.568)	(2.348)	(2.322)	(1.572)	(2.374)	(2.346)
Yield	0.721	0.942	0.970	0.487	0.655	0.666
	(0.641)	(1.040)	(1.022)	(0.573)	(0.874)	(0.866)
ΔCR	0.188	0.458	0.454	0.211	0.387	0.365
	(2.617)	(2.706)	(2.681)	(2.408)	(2.499)	(2.480)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	Yes	No	No	Yes	No
Country FE	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
N	1,184	1,184	1,184	1,184	1,184	1,184
Adj. R^2	0.414	0.403	0.411	0.412	0.402	0.410

vides a suitable estimation of a country's state of biodiversity. The variable is constructed following the same logic as in Chapter 6.3, where a binary variable is introduced and split at the median of the full available dataset. Higher scores of the measure indicate a higher vulnerability and risk exposure. They are allocated the value one and lower scores the value zero. The country split is similar compared to $BioState_{med,i}$, but not the exact same, as otherwise the results of the robustness test would not differ from those obtained in Table 13. ND-GAIN data is available yearly from 1995 until 2022. For the variables used in the analysis, the scores are averaged between 2015 and 2022. Table 18 shows the summary statistics of the scores for the full dataset. The variable can range from zero to one. In the dataset, 182 country scores are available for ecosystem vulnerability.

The regression formula for the robustness test is equal to Formula 6, where $BioState_{med,i}$ is replaced with the variable of ecosystem vulnerability, $EcoVul_{med,i}$. The findings of the regression are depicted in Table 19. The statistics in column one confirm the result of Table 13, as there is no evidence for a significant influence on the relationship between biodiversity news and CDS spread changes. The results in column two further confirm that this non-relation with a country's state of biodiversity remains, both before and after the Kunming Declaration. Note that the ND-GAIN data is not available for one country which is included in the EPI. The number of ob-

Table 18: Summary Statistics of Alternative Country Biodiversity Characteristic

This table shows the summary statistics for the country biodiversity characteristic included in the robustness check for the regression of Chapter 6.3.

Summary Statistics								
Country Factors	N	Mean	St. Dev.	Min	25th perc.	Median	75th perc.	Max
Ecosystem Vulnerability	182	0.457	0.091	0.233	0.393	0.463	0.515	0.734

Table 19: Robustness Check with alternative State of Biodiversity

This table displays the regression results of the relationship between biodiversity and CDS change depending on the state of biodiversity and the Kunming Declaration. CDS change is depicted in % change per quarter and the biodiversity measure is the standardized quarterly average of the biodiversity index. The binary variable $EcoVul_{med}$ represents the ecosystem vulnerability. Column one includes the interaction effect of the alternative state of biodiversity for the full sample and column two dependent on the Kunming Declaration. All regressions include control variables and half-yearly time fixed effects. Clustered standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)	
	(1)	(2)
Intercept	5.604**	5.624**
	(2.267)	(2.457)
Biodiversity	1.396*	0.228
	(0.736)	(1.237)
Kunming		4.514
		(3.415)
Biodiversity × Kunming		1.953
		(1.160)
$EcoVul_{med}$	-3.076***	-3.707***
	(0.738)	(0.820)
Biodiversity × $EcoVul_{med}$	0.588	0.362
	(0.867)	(1.049)
Kunming × $EcoVul_{med}$		0.410
		(1.472)
Biodiversity × Kunming × $EcoVul_{med}$		0.221
		(2.579)
Control Variables	Yes	Yes
Time FE	Yes	Yes
Bank FE	No	No
Country FE	No	No
Clustered SE	Yes	Yes
N	1,216	1,216
Adj. R^2	0.428	0.427

servations in Table 19 is therefore slightly lower at 1,216. While not reported, running the regression in Chapter 6.3 without the affected bank does not alter the results obtained in Table 13 and therefore does not interfere with the comparability of both country characteristics.

8 Conclusion

This thesis explores the relationship between biodiversity risk and bank's credit risk. Using a news-based biodiversity index by Giglio et al. (2024) for biodiversity risk and CDS as a measure of credit risk, a sample of 1,248 total observations for 39 global banks between 2015 and 2023 is constructed. This thesis finds evidence for a significant positive relationship between biodiversity news and CDS spread changes, where negative news leads to increasing CDS prices. Furthermore, the results show that equity return and volatility are dominating drivers of CDS spread changes. While balance sheet variables are shown to impact CDS price levels by similar studies, this thesis finds no evidence that this effect persists for CDS price changes.

Given the differences between US banks and non-US banks, the baseline regression will be conducted for a sub-sample. Excluding US banks will provide robustness for potential selection bias, as they constitute more than one quarter of the banks in the sample. The exclusion leads to increasing VIF estimates across all regressions. For two-way and three-way interactions, the test statistic indicates perfect multicollinearity between variables. Therefore, only the baseline regression will be tested. The results are presented

Table 20: Robustness Check for the Baseline Regression excluding US-Banks

This table provides robustness checks for the baseline regression presented in Chapter 6 both with and without time lag, by excluding US-banks. Panel A shows the results without time lag for the balance sheet variables and Panel B with time lag. Columns 1 and 4 show the regression with time fixed effects. Columns 2 and 5 show time and bank fixed effects and columns 3 and 6 time and country fixed effects. The dependent variable is the quarterly change in CDS prices. Clustered standard errors are given in parentheses and p-values are depicted as follows: (***) < 0.01, (**) < 0.05, (*) < 0.1.

	Δ CDS (%)					
	Panel A			Panel B		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	5.179**	1.124	0.595	4.527	0.626	-0.067
	(2.521)	(3.120)	(2.968)	(2.700)	(3.483)	(3.345)
Biodiversity	1.768*	1.701*	1.787*	1.726*	1.725*	1.725*
	(0.877)	(0.889)	(0.883)	(0.891)	(0.898)	(0.891)
ΔLEV	0.509	0.196	0.211	-0.397	-0.525	-0.529
	(0.864)	(0.853)	(0.845)	(0.744)	(0.754)	(0.742)
ΔAQ	0.027	0.027	0.033	-0.047	-0.055	-0.047
	(0.047)	(0.049)	(0.049)	(0.058)	(0.064)	(0.061)
ΔSize	-3.165	-1.282	-1.062	-5.204*	-3.372	-3.006
	(2.749)	(3.061)	(2.955)	(2.605)	(2.931)	(2.715)
ΔFS	0.054	0.034	0.032	-0.153	-0.173	-0.177
	(0.105)	(0.104)	(0.103)	(0.128)	(0.134)	(0.133)
ΔCE	-0.012	-0.008	-0.007	-0.006	-0.003	-0.002
	(0.038)	(0.039)	(0.039)	(0.025)	(0.026)	(0.026)
ΔMS	0.053*	0.053*	0.054*	0.008	0.007	0.008
	(0.026)	(0.028)	(0.027)	(0.018)	(0.020)	(0.019)
ΔBVOL	0.115***	0.112***	0.112***	0.114***	0.112***	0.112***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.027)	(0.027)
ΔIDXRET	-0.852***	-0.865***	-0.864***	-0.871***	-0.884***	-0.884***
	(0.100)	(0.102)	(0.101)	(0.097)	(0.101)	(0.100)
IDXVOL	6.037***	7.217***	7.225***	5.723***	6.749***	6.760***
	(1.335)	(2.026)	(2.006)	(1.489)	(2.221)	(2.192)
Yield	1.058	1.723	1.741	0.638	1.185	1.210
	(0.662)	(1.097)	(1.086)	(0.546)	(0.898)	(0.891)
ΔCR	2.947	3.334	3.326	3.087	3.266	3.236
	(2.843)	(2.892)	(2.874)	(2.591)	(2.662)	(2.642)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	Yes	No	No	Yes	No
Country FE	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
N	928	928	928	928	928	928
Adj. R^2	0.404	0.392	0.399	0.401	0.389	0.397

in Table 20. They reveal that the significance of variables from the baseline regression persists after excluding US banks. Bank volatility, as well as stock index returns and stock index volatility, are still significant at the 1% level. The biodiversity variable is significant at the 10% level instead of the 5% level seen in the baseline regression. However, as mentioned, the exclusion of US banks increased multicollinearity concerns, which may cause this change in significance through inflated standard errors. Additionally, the coefficient of *Biodiversity* is slightly higher compared to the baseline regression, as it increases from 1.532 to 1.768 in column one. This change also shows the heterogeneous effects for US and non-US banks over the sample period, as previously discussed in Chapter 6.5. Notably, for the non-US sample, market sensitivity is additionally significant at the 10% level. Overall, the implications of the results presented in Chapter 6.1 do not change when limiting the data to non-US banks, which provides evidence against a potential selection bias. Concluding, these tests have shown the robustness of the results obtained in Chapter 6.

Additionally, this thesis conducts several cross-sectional analyses to test for heterogeneous effects within the sample. The results show that the relationship persists for the period after the Kunming Declaration, suggesting no significant effect of biodiversity risk before. Furthermore, there exists no evidence for cross-sectional differences depending on a coun-

try's state of biodiversity or engagement in initiatives related to biodiversity conservation. Additional tests are undertaken to find heterogeneous results before and after the Kunming Declaration. A country's state of biodiversity does not change the relationship. In contrast, banks who engage in sustainability initiatives are found to experience lower effects of biodiversity risks in the pre-Kunming period and stronger effects afterwards. These results are conceptually in line with findings of Kölbel et al. (2022), who show a decreasing effect of climate risk on CDS before the Paris Agreement in 2015 and increasing effects afterwards, for firms who openly disclose their risk exposure. They account these effects to lower risk uncertainty before and higher risk perception effects after the external shock. This thesis further tests for heterogeneity between US and non-US banks, as the USA is the only UN nation which is not a member of the CBD. Before the Kunming Declaration, banks experienced a positive effect and afterwards a negative effect on the relationship between biodiversity risk and CDS, compared to banks located in other countries. The lower effects after the declaration are in line with the expectation of less stringent policies and regulations. The unexpected effects in the pre-Kunming period are yet unobserved to the knowledge of the author but may be explained by significant nature and climate-related disasters in the USA in the late 2010s, which had severe financial repercussions. The cross-sectional differences shown in the post-Kunming period are limited by a low number of observations, which might cause multicollinearity issues. The results are therefore left to be confirmed by similar studies. Overall, the findings of this thesis have important implications. First, banks are shown to be subject to biodiversity-related credit risks. Market participants have started to adapt CDS prices to emerging negative biodiversity news and require a higher spread since the Kunming Declaration. Therefore, banks need to consider biodiversity risks as a substantial part of their risk management processes. Furthermore, expectations of new policies and regulations following the Kunming Declaration spark the need for proactive risk management measures, as current business processes may be affected.

This thesis may provide a basis for further research. Studying the shown effects with more comprehensive samples might improve the robustness of the results. The biodiversity index by Giglio et al. (2023) currently does not provide data beyond 2023 and therefore constrains the sample of this thesis. The extension of the index would especially allow for more robust results for the period since the Kunming Declaration. Moreover, current developments, like the introduction of the TNFD framework, may be used to unveil the effects of biodiversity risk disclosure on the banking sector.

References

Ahmad, M. F., & Karpuz, A. (2024). Beyond climate change risk: Biodiversity and corporate cash holdings. *Economics Letters*, *236*. https://doi.org/10.1016/j.econlet.2024.111608

Alin, A. (2010). Multicollinearity. *WIREs Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Almeida, E., Dikau, S., Le Calvar, E., Jun, M., Kunesch, N., Nicholls, M., Robins, N., Salin, M., Sun, T., & Svartzman, R. (2022). Central banking and supervision in the biosphere: An agenda for action on biodiversity loss, financial risk and system stability. *NGFS Occasional Papers*.

Apergis, N., Danuletiu, D., & Xu, B. (2022). CDS spreads and COVID-19 pandemic. *Journal of International Financial Markets, Institutions and Money*, *76*. https://doi.org/10.1016/j.intfin.2021.101433

Bach, T. N., Hoang, K., & Le, T. (2024). Biodiversity risk and firm performance. https://doi.org/http://dx.doi.org/10.2139/ssrn.4735012

Baker, S. R., Bloom, N., Davis, S. J., & Terry, S. J. (2020). COVID-Induced Economic Uncertainty. *National Bureau of Economic Research Working Paper Series*, (26983). https://doi.org/10.3386/w26983

Barger, N. N., Gardner, T. A., Sankaran, M., Belnap, J., Broadhurst, L., Brochier, V., Isbell, F., Meyfroidt, P., Moreira, F., Nieminen, T. M., Okuro, T., Rodrigues, R. R., Saxena, V., & Ross, M. (2018). Chapter 3: Direct and indirect drivers of land degradation and restoration. In L. Montanarella, R. Scholes, & A. Brainich (Eds.), *IPBES (2018): The IPBES assessment report on land degradation and restoration* (pp. 137–218). Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. https://doi.org/10.5281/zenodo.3237393

Blanco, R., Brennan, S., & Marsh, I. W. (2005). An Empirical Analysis of the Dynamic Relation between Investment-Grade Bonds and Credit Default Swaps. *The Journal of Finance*, *60*(5), 2255–2281. https://doi.org/10.1111/j.1540-6261.2005.00798.x

Blasberg, A., Kiesel, R., & Taschini, L. (2024). Carbon default swap – disentangling the exposure to carbon risk through CDS. https://doi.org/https://dx.doi.org/10.2139/ssrn.3856993

Block, S., Emerson, J. W., Esty, D. C., de Sherbinin, A., Wendling, Z. A., et al. (2024). 2024 Environmental Performance Index.

Bolívar, F., Duran, M. A., & Lozano-Vivas, A. (2023). Business model contributions to bank profit performance: A machine learning approach. *Research in International Business and Finance*, *64*, 101870. https://doi.org/10.1016/j.ribaf.2022.101870

Boyarchenko, N., Costello, A. M., & Shachar, O. (2019). The Long and Short of It: The Post-Crisis Corporate CDS. *FRB of New York Staff Report*, (879). https://doi.org/https://dx.doi.org/10.2139/ssrn.3334640

Calice, P., Diaz Kalan, F. A., Dunz, N. M. F., & Miguel Liriano, F. (2023). Biodiversity and Finance: A Preliminary Assessment of Physical Risks for the Banking Sector in Emerging Markets. http://documents.worldbank.org/curated/en/099526305022388443

Calice, P., Kalan, F. D., & Miguel, F. (2021). Nature-Related Financial Risks in Brazil. http://documents.worldbank.org/curated/en/105041629893776228

Carbone, S., Giuzio, M., Kapadia, S., Krämer, J. S., Nyholm, K., & Vozian, K. (2021). The low-carbon transition, climate commitments and firm credit risk. *ECB Working Paper*, *2021/2631*. https://doi.org/10.2139/ssrn.3991358

CBD. (2025). List of Parties.

CBD Secretariat. (2011). Convention on Biological Diversity - Text and Annexes.

CBD Secretariat. (2021a). Financial Sector Guide for the Convention on Biological Diversity - Key Actions for Nature.

CBD Secretariat. (2021b). Kunming Declaration.

Chen, C., Ian Noble, I., Hellmann, J., Coffee, J., & Murillo, M. (2024). University of Notre Dame Global Adaptation Initiative - Country Index Technical Report. https://gain.nd.edu/assets/581554/nd_gain_countryindex_technicalreport_2024.pdf

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, *20*(2), 249–272. https://doi.org/10.1016/S0261-5606(00)00048-6

Christ, J., Hertel, T., & Kocian, J. (2022). Corporate Sustainability Performance and the Cost of Debt – An Analysis of the Impact of Country- and Industry-specific Climate Risk Exposures. https://doi.org/http://dx.doi.org/10.2139/ssrn.4135247

Coqueret, G., Giroux, T., & Zerbib, O. D. (2025). The biodiversity premium. *Ecological Economics*, *228*. https://doi.org/10.1016/j.ecolecon.2024.108435

Corò, F., Dufour, A., & Varotto, S. (2013). Credit and liquidity components of corporate CDS spreads. *Journal of Banking & Finance*, *37*(12), 5511–5525. https://doi.org/10.1016/j.jbankfin.2013.07.010

Costola, M., & Vozian, K. (2025). Pricing climate transition risk: Evidence from European corporate CDS. *Energy Economics*, *143*. https://doi.org/10.1016/j.eneco.2025.108248

Créti, A., Huang, Y., Jiang, B., & Sanin, M. E. (2024). Biodiversity Risk, Firm Performance, and Market Mispricing. https://doi.org/http://dx.doi.org/10.2139/ssrn.4765039

Cui, L., Yue, S., Nghiem, X.-H., & Duan, M. (2023). Exploring the risk and economic vulnerability of global energy supply chain interruption in the context of Russo-Ukrainian war. *Resources Policy*, *81*. https://doi.org/10.1016/j.resourpol.2023.103373

Dasgupta, P. (2021). The Economics of Biodiversity: The Dasgupta Review.

Deloitte. (2024). How banks can help achieve nature-positive outcomes and preserve biodiversity.

Drago, D., Tommaso, C. D., & Thornton, J. (2017). What determines bank CDS spreads? Evidence from European and US banks. *Finance Research Letters*, *22*, 140–145. https://doi.org/10.1016/j.frl.2016.12.035

Dudley, N. (2023). *Why Biodiversity Matters* (1st). Taylor & Francis Group. https://doi.org/10.4324/9780429346675

Einhorn, C. (2025). Far From Washington, Some Small Wins in the Fight to Protect Nature. *The New York Times*. https://www.nytimes.com/2025/02/27/climate/cop16-biodiversity-rome.html

Elderson, F. (2023). The economy and banks need nature to survive. *ECB*. https://www.ecb.europa.eu/press/blog/date/2023/html/ecb.blog230608~5cffb7c349.en.html

Ericsson, J., Jacobs, K., & Oviedo, R. (2009). The Determinants of Credit Default Swap Premia. *Journal of Financial and Quantitative Analysis*, *44*(1), 109–132. https://doi.org/10.1017/S0022109009090061

Ferreira, V., Abreu, A., & Louçã, F. (2025). The rise and fall of inflation in the Euro Area (2021-2024): A heterodox perspective. *Structural Change and Economic Dynamics*, *72*, 103–110. https://doi.org/10.1016/j.strueco.2024.12.004

FfB. (2024). Finance for Biodiversity - Reverse nature loss in this decade. https://www.financeforbiodiversity.org/wp-content/uploads/FinanceForBiodiversity_Guidance.pdf

Garel, A., Romec, A., Sautner, Z., & Wagner, A. F. (2024). Do investors care about biodiversity? *Review of Finance*, *28*(4), 1151–1186. https://doi.org/10.1093/rof/rfae010

Giglio, S., Kuchler, T., Stroebel, J., & Wang, O. (2024). The Economics of Biodiversity Loss. https://doi.org/10.3386/w32678

Giglio, S., Kuchler, T., Stroebel, J., & Zeng, X. (2023). Biodiversity Risk. https://doi.org/10.3386/w31137

Hasan, I., Liu, L., & Zhang, G. (2016). The Determinants of Global Bank Credit-Default-Swap Spreads. *Journal of Financial Services Research*, *50*(3), 275–309. https://doi.org/10.1007/s10693-015-0232-z

Hasan, I., Marra, M., To, T. Y., Wu, E., & Zhang, G. (2023). COVID-19 Pandemic and Global Corporate CDS Spreads. *Journal of Banking & Finance*, *147*. https://doi.org/10.1016/j.jbankfin.2022.106618

Hoepner, A. G. F., Klausmann, J., Leippold, M., & Rillaerts, J. (2023). Beyond Climate: The Impact of Biodiversity, Water, and Pollution on the CDS Term Structure [Swiss Finance Institute Research Paper No. 23-10, Michael J. Brennan Irish Finance Working Paper Series Research Paper No. 23-4]. https://doi.org/http://dx.doi.org/10.2139/ssrn.4351633

Hudson, R. (2024). Biodiversity and risk in the financial sector. *Journal of Sustainable Finance and Accounting*, *2*. https://doi.org/10.1016/j.josfa.2024.100009

IHS Markit. (2021). CDS Indices Primer.

IPBES. (2019). Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

Jorion, P., & Zhang, G. (2007). Good and bad credit contagion: Evidence from credit default swaps. *Journal of Financial Economics*, *84*(3), 860–883. https://doi.org/10.1016/j.jfineco.2006.06.001

Kalhoro, M. R., & Kyaw, K. (2024). Manage biodiversity risk exposure? *Finance Research Letters*, *61*. https://doi.org/10.1016/j.frl.2024.104989

Karolyi, G. A., & Tobin-de la Puente, J. (2023). Biodiversity finance: A call for research into financing nature. *Financial Management*, *52*(2), 231–251. https://doi.org/10.1111/fima.12417

Kölbel, J., Leippold, M., Rillaerts, J., & Wang, Q. (2022). Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks affects the CDS Term Structure. Swiss Finance Institute Research Paper No. 21-19. https://doi.org/http://dx.doi.org/10.2139/ssrn.3616324

Lee, J., Naranjo, A., & Velioglu, G. (2018). When do CDS spreads lead? Rating events, private entities, and firm-specific information flows. *Journal of Financial Economics*, *130*(3), 556–578. https://doi.org/10.1016/j.jfineco.2018.07.011

Liang, C., Yang, J., Shen, L., & Dong, D. (2024). The role of biodiversity risk in stock price crashes. *Finance Research Letters*, *67*. https://doi.org/10.1016/j.frl.2024.105856

Longstaff, F. A., Mithal, S., & Neis, E. (2005). Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from the Credit Default Swap Market. *The Journal of Finance*, *60*(5), 2213–2253. https://doi.org/10.1111/j.1540-6261.2005.00797.x

McKinley, D. C., Ryan, M. G., Birdsey, R. A., Giardina, C. P., Harmon, M. E., Heath, L. S., Houghton, R. A., Jackson, R. B., Morrison, J. F., Murray, B. C., Pataki, D. E., & Skog, K. E. (2011). A synthesis of current knowledge on forests and carbon storage in the United States. *Ecological Applications*, *21*(6), 1902–1924. https://doi.org/10.1890/10-0697.1

MSCI ESG Research LLC. (2023). MSCI ESG Ratings Methodology: Biodiversity & Land Use Key Issue.

Naffa, H., & Czupy, G. (2024). Biodiversity Risk Premium.

Norden, L., & Weber, M. (2009). The Co-movement of Credit Default Swap, Bond and Stock Markets: an Empirical Analysis. *European Financial Management*, *15*(3), 529–562. https://doi.org/10.1111/j.1468-036X.2007.00427.x

Noth, F., & Schüwer, U. (2023). Natural disasters and bank stability: Evidence from the U.S. financial system. *Journal of Environmental Economics and Management*, *119*. https://doi.org/10.1016/j.jeem.2023.102792

OECD. (2023a). A supervisory framework for assessing nature-related financial risks: Identifying and navigating biodiversity risks. *OECD Business and Finance Policy Papers*, *33*. https://doi.org/10.1787/a8e4991f-en

OECD. (2023b). Assessing biodiversity-related financial risks: Navigating the landscape of existing approaches. https://doi.org/10.1787/d52137a5-en

Pörtner, H. O., Scholes, R. J., Agard, J., Archer, E., Arneth, A., Bai, X., Barnes, D., Burrows, M., Chan, L., Cheung, W. L., Diamond, S., Donatti, C., Duarte, C., Eisenhauer, N., Foden, W., Gasalla, M. A., Handa, C., Hickler, T., Hoegh-Guldberg, O., & Ngo, H. T. (2021). Scientific outcome of the IPBES-IPCC co-sponsored workshop on biodiversity and climate change. B. IPBES secretariat, Germany. https://doi.org/10.5281/zenodo.4659158

Refinitiv Datastream. (n.d.). Refinitiv Datastream - The world's most comprehensive financial historical database.

Smith, A. B. (2019). 2018's Billion Dollar Disasters in Context. https://www.climate.gov/news-features/blogs/beyond-data/2018s-billion-dollar-disasters-context

Smith, A. B. (2020). 2010-2019: A landmark decade of U.S. billion-dollar weather and climate disasters. https://www.climate.gov/news-features/blogs/beyond-data/2010-2019-landmark-decade-us-billion-dollar-weather-and-climate

Smith, K. W., & Sasaki, M. S. (1979). Decreasing Multicollinearity: A Method for Models with Multiplicative Functions. *Sociological Methods & Research*, *8*(1), 35–56. https://doi.org/10.1177/004912417900800102

Stocker, M., Baffes, J., Some, Y. M., Vorisek, D., & Wheeler, C. M. (2018). The 2014-16 Oil Price Collapse in Retrospect: Sources and Implications. *Policy Research Working Paper*, (8419). http://hdl.handle.net/10986/29756

Svartzman, R., Espagne, E., Julien, G., Paul, H.-L., Mathilde, S., Allen, T., Berger, J., Calas, J., Godin, A., & Vallier, A. (2021). A 'Silent Spring' for the Financial System? Exploring Biodiversity-Related Financial Risks in France. *Banque de France Working Paper*, (826). https://doi.org/http://dx.doi.org/10.2139/ssrn.4028442

Tang, D. Y., & Yan, H. (2010). Market conditions, default risk and credit spreads. *Journal of Banking & Finance*, *34*(4), 743–753. https://doi.org/10.1016/j.jbankfin.2009.05.018

Tate, R. L. (1984). Limitations of Centering for Interactive Models. *Sociological Methods & Research*, *13*(2), 251–271. https://doi.org/10.1177/0049124184013002004

TNFD. (2023). Recommendations of the Taskforce on Nature-related Financial Disclosures.

Ugolini, A., Reboredo, J. C., & Ojea-Ferreiro, J. (2024). Is climate transition risk priced into corporate credit risk? Evidence from credit default swaps. *Research in International Business and Finance*, *70*. https://doi.org/10.1016/j.ribaf.2024.102372

UNEP FI. (2022). Principles for Responsible Banking - Guidance Document.

UNEP FI. (2024). 2023 Annual Overview.

van Toor, J., Piljic, D., Schellekens, G., van Oorschot, M., & Kok, M. (2020). Indebted to nature: Exploring biodiversity risks for the Dutch financial sector. https://www.dnb.nl/media/4c3fqawd/indebted-to-nature.pdf

Wang, D., van Lelyveld, I., & Schaumburg, J. (2019). Do Information Contagion and Business Model Similarities Explain Bank Credit Risk Commonalities? *ESRB: Working Paper Series*, (2019/94). http://dx.doi.org/10.2139/ssrn.3723456

WEF. (2020). Nature Risk Rising: Why the Crisis Engulfing Nature Matters for Business and the Economy. https://www3.weforum.org/docs/WEF_New_Nature_Economy_Report_2020.pdf

WEF. (2024). The Global Risks Report 2024 (19th Edition). https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Wooldridge, J. M. (2020). *Introductory econometrics a modern approach* (Seventh). Cengage.

World Bank Group. (2016). Global Economic Prospects: Divergences and Risks. http://documents.worldbank.org/curated/en/588301469460261009

Yin, J., Han, B., & Wong, H. Y. (2022). COVID-19 and credit risk: A long memory perspective. *Insurance: Mathematics and Economics*, *104*, 15–34. https://doi.org/10.1016/j.insmatheco.2022.01.008

Zhang, B. Y., Zhou, H., & Zhu, H. (2009). Explaining Credit Default Swap Spreads with the Equity Volatility and Jump Risks of Individual Firms. *The Review of Financial Studies*, *22*(12), 5099–5131. http://www.jstor.org/stable/40468339

Zhang, Z., & Zhao, R. (2022). Carbon emission and credit default swaps. *Finance Research Letters*, *50*. https://doi.org/10.1016/j.frl.2022.103286

Zhu, H. (2006). An Empirical Comparison of Credit Spreads between the Bond Market and the Credit Default Swap Market. *Journal of Financial Services Research*, *29*(3), 211–235. https://doi.org/10.1007/s10693-006-7626-x

List of Abbreviations

bp Basis Points

CBD Convention on Biological Diversity

CBF Corporate Biodiversity Footprint

CDS Credit Default Swaps

COP 15 15th Conference of the Parties to the Convention on Biological Diversity

CR Full Restructuring

ECB European Central Bank

EPI Environmental Performance Index

FfB Finance for Biodiversity

IPBES Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services

ISDA International Swaps and Derivatives Association

MM Modified-Modified Restructuring

MR Modified Restructuring

ND-GAIN Notre Dame Global Adaptation Initiative

NOAA National Oceanic and Atmospheric Administration

OECD Organisation for Economic Co-operation and Development

OLS Ordinary Least Squares

TNFD Task Force on Nature-related Financial Disclosures

UN United Nations

UNEP FI United Nations Environment Program Finance Initiative

VIF Variance Inflation Factor

WEF World Economic Forum

XR No Restructuring

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Kausale Inferenz unter Anwendung von Double Machine Learning: Oregon Health Insurance Experiment

Sebastian Schmidt

University of Hamburg

Abstract

This study applies advanced methods of causal inference, specifically the Double Machine Learning (DML) framework, to estimate the causal effects of public health insurance on individual health outcomes using data from the 2008 Oregon Health Insurance Experiment (OHIE), a randomized controlled trial. DML integrates modern machine learning with the econometric principles of causal identification to obtain unbiased treatment effect estimates in high-dimensional data. Interactive Regression Models (IRM) and Interactive Instrumental Variable Models (IIVM) are employed to estimate effects of Medicaid coverage on perceived health, number of doctor visits, satisfaction, access to medical services, and quality of care. The results indicate small but positive causal effects of Medicaid coverage on perceived health and healthcare utilization, statistically insignificant effects on satisfaction and access to medication, and a slightly negative effect on perceived quality of care. The findings highlight the potential of Double Machine Learning as a robust framework for causal analysis in empirical research.

Zusammenfassung

Diese Arbeit verwendet fortgeschrittene Methoden der Kausalen Inferenz, insbesondere das Double Machine Learning (DML) Framework, um die kausalen Effekte f́6ffentlicher Krankenversicherung auf individuelle Gesundheitsvariablen anhand der Daten des Oregon Health Insurance Experiments (OHIE, 2008) zu sché4tzen. DML ermf́6glicht, unverzerrte Sché4tzungen von Treatment-Effekten in hochdimensionalen Datensé4tzen zu ermf́6glichen. Mittels Interactive Regression Models (IRM) und Interactive Instrumental Variable Models (IIVM) werden Effekte von Medicaid auf wahrgenommene Gesundheit, Anzahl der Arztbesuche, Zufriedenheit, Zugang zu medizinischen Leistungen sowie Versorgungsqualité4t untersucht. Die Ergebnisse zeigen geringe, aber positive kausale Effekte von Medicaid auf die wahrgenommene Gesundheit und die Hé4ufigkeit der Arztbesuche, statistisch insignifikante Effekte auf Zufriedenheit und Medikamentenzugang sowie einen leicht negativen Zusammenhang mit der wahrgenommenen Behandlungsqualité4t. Die Ergebnisse unterstreichen das Potenzial von Double Machine Learning als robustes Instrument der Kausalanalyse in empirischen Studien.

Keywords: causal inference; double machine learning; oregon health insurance experiment

1 Einleitung

In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist. Englische Fachbegriffe ohne geläufiges deutsches Äquivalent werden im Original verwendet, da diese im Fachbereich gängig sind.

Eine der grundlegenden Fragen der Wissenschaft lautet „Warum?“. Seit Jahrtausenden treibt das Bestreben nach einer Erklärung für die beobachteten Phänomene der Welt Wissenschaftler sowie neugierige Denker an. Dabei kann diese Frage unzählig viele Facetten annehmen: Warum gewittert es? Warum fallen Äpfel auf den Boden? Warum werden Menschen krank? All das und viele weitere, sind Fragen nach der Ursächlichkeit bzw. der Kausalität. Es stellt sich die Frage, weshalb Kausalität so interessant für die Menschen ist. Und die Antwort liegt nahe: Das Verständnis von kausalen Zusammenhängen ermöglicht uns, Kausaleffekte zu reproduzieren, den Verlauf von Ereignissen in der Zukunft zu manipulieren und mit alledem zu unseren Gunsten zu wenden. Dieses Verständnis kann auf einer sehr oberflächlichen oder einer fundamentalen Ebene vorliegen - in beiden Fällen kann ein Mehrwert aus dem Erkennen kausaler Zusammenhänge geschlossen werden. „Das Essen dieser Pflanze führt zu Bauchschmerzen. Ich sollte diese Pflanze nicht mehr essen“ oder „Das Drehen einer metallischen Spule innerhalb eines magnetischen Felds induziert Elektrizität, die wir nutzen können, um elektrische Geräte mit Energie zu versorgen“. Das Wissen, dass das Eintreten von A anschließend das Eintreten von B auslöst oder begünstigt, bietet einen Mehrwert. Der Mensch neigt von Natur aus dazu, kausale Zusammenhänge schnell und intuitiv zu erlernen. Dennoch könnte es in

DOI: https://doi.org/10.5282/jums/v11i1pp107-138

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

der Realität kaum schwieriger sein, kausale Zusammenhänge zu beweisen. Ist der Patient einer Studie tatsächlich genesen, weil er einer neuen Behandlungsmethode unterzogen wurde oder war es Zufall und der Patient wäre auch ohne medizinisches Eingreifen wieder gesund geworden?

Das akademische Feld, das sich damit beschäftigt Kausalität zu untersuchen, ist die Kausale Inferenz. Kausalität und damit auch Kausale Inferenz haben eine große Relevanz in vielen Bereichen des alltäglichen Lebens. Wie am zuvor genannten Beispiel ersichtlich, ist Kausalität enorm wichtig in der Medizin. Moderne Medizin baut generell darauf auf, dass der Behandlung einer Krankheit ein kausaler Effekt zur Genesung unterstellt wird - andernfalls gäbe es keinen Grund für die Anwendung einer Behandlung (der Placebo-Effekt ist dabei eingeschlossen, da auch dieser kausal ausgelöst werden kann und zur Genesung eines Patienten beiträgt) (Thompson & Upshur, 2017). Wäre Wissenschaftlern und Ärzten der volle kausale Zusammenhang zwischen sämtlichen Medikamenten, Behandlungen und Erkrankungen klar, so wären viele Krankheiten heilbar. Jedoch sind kausale Zusammenhänge häufig komplizierter als nur die Verabreichung eines Medikaments. Bei vielen Krankheiten kommt es beispielsweise auch darauf an, in welchem Stadium der Erkrankung eine Behandlung angewendet wird - in der Regel gilt hier, je früher die Therapie, desto höher die Chance auf Erfolg. Damit eine Krankheit rechtzeitig behandelt werden kann, muss sie auch rechtzeitig erkannt werden (Ott et al., 2009). Meist ist der Besuch eines Krankenhauses oder Arztes bei Auftreten erster Symptome dabei der entscheidende Faktor. Abhängig vom Gesundheitssystem des jeweiligen Landes, in dem potenzielle Patienten sich befinden, funktioniert der Besuch eines Arztes jedoch unterschiedlich. Einer der größten Unterschiede ist dabei die Abgrenzung zwischen privater und öffentlicher Gesundheitsversorgung. In Ländern wie Deutschland und Schweden sind die Einwohner standardmäßig über die staatliche öffentliche Krankenkasse versichert und können abgesehen von einem monatlichen Beitrag alle von der Krankenkasse anerkannten Leistungen kostenfrei oder unter geringer Kostenbeteiligung in Anspruch nehmen (§5 und §6 (*SGB V*, 1989)). Besuche bei niedergelassenen Ärzten und Krankenhäusern sind dabei gleichermaßen inbegriffen. In den USA hingegen ist das Gesundheitssystem weitgehend privatisiert. Dort gibt es kein öffentliches Gesundheitssystem, welches standardmäßig jeden Einwohner abdeckt. Der Default in den USA ist es, nicht krankenversichert zu sein und entweder eine private Krankenversicherung abzuschließen oder im Falle einer medizinischen Untersuchung oder Behandlung selbst für die Kosten aufzukommen (Vladeck, 2003). Eine Problematik bei diesem System ist, dass ein großer Anteil der Bevölkerung der USA sich den Abschluss einer privaten Krankenversicherung finanziell nicht leisten kann (Dickman et al., 2017). Generell sind medizinische Leistungen teuer - ein einfaches Röntgenbild kann tausende Dollar kosten, in Ausnahmefällen sogar mehr. Eine Chemotherapie zur Behandlung einer Krebserkrankung erreicht schnell den Wert eines Einfamilienhauses bis hin zu Millionenbeträgen (Kantarjian & Rajkumar, 2015). Das sind Kosten, die selbst wohlhabende Menschen schnell finanziell gefährden könnten. Dieser Umstand lässt Raum für Vermutung, dass die Art der Gesundheitsversicherung eine Reihe an Faktoren beeinflussen kann. Es wäre denkbar, dass Menschen ihr Verhalten daran anpassen, ob sie die Kosten für medizinische Leistungen selbst tragen müssen. Beispielsweise könnten die Anzahl an Arztbesuchen innerhalb eines Intervalls sowie der physische und psychische Gesundheitszustand der Bevölkerung und ähnliche Variablen von der Art der Versicherung beeinflusst sein. Der Vorteil von einem öffentlichen Gesundheitssystem ist, dass die Kosten für medizinische Behandlungen weitestgehend abgefangen werden. Darauf lässt sich die folgende Frage formulieren: „Welche direkten und indirekten kausalen Effekte hat Public Healthcare auf die abgedeckte Bevölkerung und wie stark sind diese?“. Bei dem Oregon Health Insurance Experiment (im Folgenden auch OHIE) aus dem Jahr 2008 wurden Daten zu genau der Beantwortung dieser Frage erhoben. Eine Auswahl an Personen aus dem Staat Oregon erhielten über einen Zeitraum von einem Jahr „MedicAid“ - also eine staatlich geförderte Gesundheitsversicherung, für die die Empfänger keine Kosten tragen mussten. Im Rahmen von mehreren Umfragen wurden im Verlauf des Experiments Daten zu den Teilnehmern erhoben. Die im OHIE gesammelten Daten wurden bereits mit traditionellen statistischen Verfahren analysiert (Baicker & Finkelstein, 2014).

Im Feld der Kausalen Inferenz sind innerhalb der letzten Jahre neue Ansätze und Methoden entwickelt worden, um Kausalität zu untersuchen. Diese Methoden sind vielversprechend, da sie einen neuen Blickwinkel auf die Daten aus dem OHIE und die assoziierten Fragen eröffnen. Eine dieser Methoden ist das Double Machine Learning Framework, welches versucht mithilfe von Machine Learning Algorithmen kausale Effekte zu berechnen (Chernozhukov et al., 2016). Seit Veröffentlichung des OHIE wurden bereits Studien und Untersuchungen mit den Daten aus dem Experiment erhoben, jedoch hat keine der wissenschaftlichen Arbeiten Double Machine Learning angewandt, um die potenziell zugrundeliegenden kausalen Effekte zu untersuchen. Im Rahmen dieser Arbeit soll genau das geschehen: Aus dem Blickwinkel der Kausalen Inferenz sollen die Daten des Oregon Health Insurance Experiment unter Verwendung vom Double Machine Learning Framework auf Kausalität untersucht werden. Dabei ist die Zielsetzung dieser Arbeit, die folgenden Leitfragen zu beantworten:

- Inwieweit hat MedicAid einen Einfluss auf die Gesundheit der Probanden?
- Ist die Anzahl der Arztbesuche innerhalb eines Zeitintervalls abhängig von MedicAid-Coverage?
- Beeinflusst MedicAid die Zufriedenheit der Probanden?

- Werden ohne MedicAid-Coverage eigentlich notwendige medizinische Leistungen nicht in Anspruch genommen?
- Hat MedicAid-Coverage einen Einfluss auf die Qualität der erhaltenen medizinischen Leistungen?

Die Arbeit ist dabei so strukturiert, dass zuerst die wissenschaftlichen Grundlagen und Definitionen dargelegt werden. Dafür wird im zweiten Kapitel der Fachbereich der Kausalen Inferenz näher erläutert. Begriffe wie Korrelation und Kausalität werden sinngemäß und mathematisch eingeführt. Zudem wird das Double Machine Learning Framework in seiner Notwendigkeit und seiner Funktionsweise erklärt. Im dritten Kapitel wird der in dieser Arbeit verwendete Datensatz des Oregon Health Insurance Experiments näher vorgestellt. Dabei wird zunächst die Durchführung des Experiments und anschließend der Datensatz beschrieben. Im vierten Kapitel wird die Methodik dieser Arbeit ausführlich aufgezeigt - zum einen wird dabei auf die konzeptionelle Vorgehensweise und zum anderen auf die technische Umsetzung eingegangen. Im fünften Kapitel werden die Ergebnisse der Anwendung des Double Machine Learning Frameworks ausgewertet. Mögliche Kritikpunkte an dem verwendeten Datensatz oder der angewandten Methodik werden im sechsten Kapitel diskutiert. Anschließend werden im siebten Kapitel die Ergebnisse zusammengefasst und es wird deren Validität bewertet. Im letzten Kapitel wird ein Ausblick für zukünftige Anwendung von Double Machine Learning - insbesondere im Kontext vom OHIE - gegeben.

2 Kausale Inferenz

Kausale Inferenz ist eine moderne Weiterentwicklung der Statistik. Jahrzehntelang wurden mithilfe von statistischen Methoden deskriptiv Wahrscheinlichkeiten, Zusammenhänge und Abhängigkeiten beobachtet und erforscht. Statistik selbst ist eine akademische Disziplin, die sich der Mathematik bedient und sich mit der Sammlung, Analyse, Interpretation, Präsentation und Organisation von Daten beschäftigt (*Oxford Dictionary of English*, 2010). Sie befasst sich damit, wie Daten gesammelt und so analysiert werden können, dass sich daraus sinnvolle Schlüsse ziehen lassen. Dabei geht es darum, Muster in den Daten zu identifizieren, Zusammenhänge zwischen Variablen aufzudecken und Hypothesen auf ihre Gültigkeit zu überprüfen (Romeijn, 2022). Doch ein entscheidender Punkt, der in der klassischen Statistik schwierig abzubilden ist, ist die Wirkungsrichtung von Ereignissen. Statistische Maße können beschreiben, dass zwei Ereignisse zusammenhängen - doch sie können nicht formulieren, inwieweit sie sich eventuell gegenseitig beeinflussen (Pearl, 2009a). Das Kapitel ist wie folgt strukturiert: Das erste Unterkapitel widmet sich den zum Verständnis dieser Arbeit notwendigen Definitionen. Im zweiten Unterkapitel wird das Double Machine Learning, ein Framework zur Anwendung von Machine Learning bei der Untersuchung kausaler Effekte, erklärt. Darin wird die Bedeutung des Frameworks für die Anwendung von Machine Learning sowie sein Aufbau erklärt.

2.1 Definitionen

Nachdem zuvor bereits die wissenschaftliche Disziplin der Kausalen Inferenz eingeführt wurde, ist es zunächst wichtig, die darin verwendeten Begrifflichkeiten zu definieren. Dabei sollten die Begriffe sowohl sinngemäß, als auch mathematisch abgegrenzt werden. Gerade Letzteres ist im Bereich der Statistik, Mathematik und der Kausalen Inferenz von großer Wichtigkeit. Um Berechnungen durchführen zu können, sollte zunächst ein Grundgerüst aus Definitionen und Formeln gebaut werden, um alle anschließenden Berechnungen zu stützen. In den folgenden Abschnitten soll es darum gehen, die relevanten Begrifflichkeiten der Kausalen Inferenz zu definieren und in Formeln abzubilden. Kernbegriff im Feld der Kausalen Inferenz ist natürlich die Kausalität. Diese soll - insbesondere in Abgrenzung zur Korrelation - definiert werden. Dafür geht es im ersten Unterkapitel um Korrelationen, deren Interpretation und eine Möglichkeit, diese abzubilden. Das zweite Unterkapitel grenzt daraufhin den Begriff der Kausalität ab und gibt jeweils die Formel für den individuellen und den durchschnittlichen kausalen Effekt. Zuletzt wird Confounding, eine potenzielle Störquelle bei der Bestimmung kausaler Effekte, erklärt und definiert.

2.1.1 Korrelation

Korrelationen sind in der Statistik von großer Bedeutung. Soll die Beziehung zwischen zwei oder mehr Variablen untersucht werden, ist die Korrelation meist das erste Werkzeug (Croxton & Cowden, 1939). Es gibt verschiedene Arten Korrelationen zu messen - abhängig von den betrachteten Variablen und der gestellten Frage können verschiedene Korrelationskoeffizienten verwendet werden. Der vermutlich gängigste ist der Pearson Korrelationskoeffizient. Dieser beschreibt die Stärke des linearen Zusammenhangs zwischen zwei Variablen. Der Koeffizient bewegt sich zwischen +1 und -1, wobei 1 eine perfekte positive Korrelation, -1 eine perfekte positive Korrelation und 0 keine Korrelation bedeutet. Der Pearson-Korrelationskoeffizient kann symbolisch dargestellt werden als:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

wobei r_{xy} der Pearson-Korrelationskoeffizient zwischen den beiden Variablen x und y ist, n die Stichprobengröße, x_i und y_i die Werte der beiden Variablen für jede Beobachtung i sind und $\bar{x}$ und $\bar{y}$ ihre jeweiligen Stichprobenmittelwerte sind (Wright, 1921). Während Korrelation nützlich ist, um die Beziehung zwischen zwei Variablen zu untersuchen ist jedoch wichtig zu beachten, dass eine Korrelation allein nicht zwangsläufig eine Kausalität zwischen den betrachteten Variablen impliziert (Wright, 1921). Während der Korrelation zwischen Rauchen und Lungenkrebs eine direkte Kausalbeziehung zugrundeliegt, ist dies bei der Korrelation

zwischen der Mortalität in Krankenhäusern und dem Volumen von Operationen anders. Im Jahr 2002 fanden Birkmeyer et al., dass die Mortalität für 14 Arten von Operationen invers mit der Häufigkeit der Operationen im Krankenhaus korreliert waren (Birkmeyer et al., 2002). Diese Korrelation ist jedoch unwahrscheinlich auf eine direkte kausale Beziehung zwischen der Häufigkeit der Operationen und der Mortalität der Patienten zurückzuführen. Stattdessen könnte sie auf andere Faktoren zurückzuführen sein, wie z.B. Erfahrung des Personals, die Ausstattung oder Ressourcenallokation des Krankenhauses. Wird eine Operation beispielsweise häufig durchgeführt, so sammeln die behandelnden Ärzte Erfahrung, welche sich positiv auf den Erfolg zukünftiger Eingriffe auswirken kann. Die Korrelation kann in diesem Fall also ein Indikator für Kausalität sein, jedoch stehen nicht alle korrelierten Variablen zwangsläufig in einer direkten Kausalbeziehung zueinander. Es besteht auch die Möglichkeit, dass Korrelationen lediglich zufällig Auftreten und gar keine Kausalität zugrunde liegt. Wie britischer Ökonom und Nobelpreisträger Ronald Coase bereits sagte: „If you torture the data long enough, it will confess to anything.“ (Coase, n. d.). Diese Aussage legt nahe, dass bei ausreichend langer und intensiver Datenanalyse fast jedes gewünschte Muster oder jede Verbindung gefunden werden kann, selbst wenn diese reinem Zufall geschuldet sind. Die Schlussfolgerung daraus ist, dass bei der Interpretation von statistischen Zusammenhängen stets versucht werden sollte, die Ergebnisse durch unabhängige Replikation und Validierung zu bestätigen (Roberts et al., 2006).

Neben der gerade erläuterten Problematik gibt es eine weitere Schwierigkeit beim Arbeiten mit Korrelationen: Korrelationskoeffizienten können keine Wirkungsrichtung abbilden (Pearl, 2009a). Um dies anhand eines Beispiels zu verdeutlichen, lässt sich der folgende Fall betrachten: Kopfschmerzen und die Einnahme von Schmerzmitteln sind positiv miteinander korreliert. In diesem Fall liegt auch eine kausale Beziehung vor. Ohne ein tieferes Verständnis für den Sachverhalt besteht jedoch keine Möglichkeit anhand des Korrelationskoeffizienten eine Aussage darüber zu treffen, welche der beiden Variablen die andere „auslöst“. Mit mehr Wissen über den Sachverhalt oder einer Untersuchung der kausalen Koeffizienten wird ersichtlich, dass Kopfschmerzen begünstigen Schmerzmittel einzunehmen. Aufgrund der in diesem Abschnitt erläuterten Probleme bei der Verwendung von Korrelation zur Untersuchung kausaler Effekte, soll im folgenden Abschnitt Kausalität näher erläutert werden.

2.1.2 Kausalität und kausale Effekte

Die erste Lektion in jeder Statistik-Vorlesung und des vorigen Abschnitts ist die folgende: Korrelation bedeutet nicht gleich Kausalität. Kausalität bedeutet vereinfacht: Ein Ereignis beeinflusst oder verursacht ein anderes (Pearl, 2009a). Fällt ein Gärtner einen Baum, kippt dieser um. Das Bearbeiten des Baumes mit einer Axt oder Säge hat den Baum zum Umstürzen gebracht - somit war es der Auslöser. Es kann hier also von Kausalität gesprochen werden. Zumindest vermeintlich. Auch wenn es offensichtlich erscheint, dass der Baum lediglich umgefallen ist, weil der Gärtner ihn abgesägt hat, wäre es jedoch auch denkbar, dass der Baum im gleichen Moment ohne das Zutun des Gärtners auch umgefallen wäre. Stimmte dies, so wäre der Gärtner nicht der kausale Auslöser für das Fallen des Baumes. Kausalität liegt nämlich erst dann vor, wenn ein Ereignis A ein anderes Ereignis B auslöst und B ohne das Eintreten von A nicht passiert wäre (Pearl, 2009a). Das Ergebnis, auch bekannt als „Outcome“, sollte abhängig davon unterschiedlich sein, ob ein Eingriff (in der Kausalen Inferenz „Treatment“ genannt) durchgeführt wurde oder nicht. In der realen Welt stellt dies jedoch ein Problem dar: Entweder ein Ereignis tritt ein, oder nicht. Es kann entschieden werden, dass ein Treatment durchgeführt oder verabreicht wird - danach kann aber nicht mehr der Fall beobachtet werden, in dem das Treatment nicht verabreicht wurde. Wird ein Patient einer Operation unterzogen, so ist nach der Operation unmöglich zu erfahren, wie der Zustand desselben Patienten ohne den Eingriff gewesen wäre. Der in der Realität eingetretene Fall wird in der Kausalen Inferenz als „Factual“ bezeichnet, während der gegensätzliche Fall das sogenannte „Counterfactual“ ist (Pearl, 2009b). Es ist faktisch unmöglich beide Varianten zu beobachten. Ein Counterfactual ist per Definition nicht zu beobachten, was auch als das fundamentale Problem der Kausalen Inferenz bekannt ist (Holland, 1986). Obwohl Kausalität in der realen Welt nicht zu beweisen ist, kann sie dennoch untersucht werden. Dafür werden statistische Methoden zu Hilfe gezogen. Die bisher genannten Beispiele waren allesamt individuelle kausale Effekte - „Macht es einen Unterschied, ob wir diese Person mit einem Medikament behandeln, oder nicht?“. In der Kausalen Inferenz werden durchschnittliche kausale Effekte untersucht - „Verabreichen wir dieses Medikament einer Gruppe von Erkrankten, haben diese dann eine höhere Chance auf Genesung, als die Kontrollgruppe?“ (Pearl, 2009b).

Ist das Ziel kausale Effekte aus mathematischer Perspektive zu betrachten, so sind zunächst einige Festlegungen zu treffen. Diese wurden 1980 von Rubin unter dem Namen „Stable-Unit-Treatment-Value Assumption“, im Folgenden als SUTVA abgekürzt, festgehalten (Granger, 1986; Rubin, 1980). SUTVA lässt sich im Wesentlichen auf zwei Aussagen reduzieren:

- Es darf keine Interferenzen zwischen den Probanden geben.
- Es darf keine verschiedenen Versionen des Treatments geben.

In beiden Fällen wäre Y_i^a nicht ausreichend definiert, um kausale Effekte zu bestimmen (Rubin, 1980). Erstere Aussage bedeutet, dass der Outcome des Treatments bei einem Probanden keinen Einfluss auf den Outcome eines anderen Probanden haben darf. Werden beispielsweise zwei Geschwister einer Herzoperation unterzogen, wäre es denkbar, dass die erfolgreiche Operation bei der ersten Person und die damit einhergehende Erleichterung bei Proband und Arzt einen positiven Effekt auf das Ergebnis der zweiten Operation hätten

(oder vice versa) und somit den kausalen Effekt verfälscht. Am selben Beispiel lässt sich auch die zweite Festlegung erklären. Es darf keine verschiedenen Versionen des Treatments geben: Das heißt, es muss für eine Vergleichbarkeit die exakt selbe Operation an allen Probanden durchgeführt werden. Wird bei Proband *A* ein Herzschrittmacher eingesetzt und bei Proband *B* wird ein Bypass gelegt, so sind die Eingriffe nicht vergleichbar miteinander. Patienten denen ein Herzschrittmacher eingesetzt wurde, sind jedoch untereinander vergleichbar. In der tiefsten Bedeutung dieser Annahme steckt, dass Treatments zu 100% übereinstimmen müssten, bis ins kleinste Detail. Unterscheidet sich der behandelnde Arzt, müsste von verschiedenen Treatments gesprochen werden. Streng genommen würde sich das Treatment bereits unterscheiden, wenn eine Operation 5 Sekunden länger dauert als eine andere oder der behandelnde Arzt bereits einen langen Arbeitstag hinter sich hat (und bei einer anderen Operation nicht). Es mag dem praktischen Nutzen nicht dienen, die Definitionen so streng zu befolgen, aber jegliche Abmilderung der Annahmen sind potenziell Fehlerquellen bei der Ermittlung von Kausalität (Rubin, 1980). In der Regel ist trotz bester Intentionen eine so strenge Durchführung jedoch nicht möglich und es wird versucht, Treatments so vergleichbar wie realistischerweise möglich zu halten. Im Nachhinein kann dann bewertet werden, ob von einem konstanten Treatment gesprochen werden kann bzw. ob die Variationen im Treatment irrelevant für die Bestimmung des kausalen Effekts sind. Im Fall des Oregon Health Insurance Experiments sind die SUTVA-Annahmen jedoch weitgehend erfüllt. Das in dieser Arbeit betrachtete Treatment ist eine Aufnahme in das MedicAid-Programm. Das MedicAid-Programm und die dazugehörigen Leistungen sind für alle Teilnehmer identisch und die Aufnahme in das Programm hat keinen Einfluss darauf, ob eine andere Person ebenfalls aufgenommen wird oder nicht.

Nachdem die wesentlichen Annahmen getroffen sind, kann begonnen werden, kausale Effekte durch mathematische Formeln zu definieren: Es gibt eine dichotome bzw. binäre Treatment Variable, im Folgenden *A* genannt; $A = 1$ bedeutet, das Treatment wurde angewandt, $A = 0$ steht für unbehandelt. Das Wort „Treatment“ kann dabei auch auch außerhalb des medizinischen Kontextes verstanden werden. Das Ausstrahlen eines Werbespots kann ebenfalls ein Treatment sein. Der Outcome wird in der Variable *Y* festgehalten, welche in diesem Fall ebenfalls dichotom ist. Geht es um die Wirksamkeit eines neuen Medikaments, so ergäbe sich daraus beispielsweise folgende Übersicht:

A = 0: Das Medikament wird nicht verabreicht
A = 1: Das Medikament wird verabreicht
Y = 0: Der Patient wird nicht geheilt
Y = 1: Der Patient wird geheilt

Daraus lässt sich die Formel für den individuellen kausalen Effekt wie folgt formulieren:

Das Treatment *A* hat einen kausalen Effekt auf den Outcome *Y*, wenn

$$Y^{a=1} \neq Y^{a=0} \tag{2}$$

(Hernan & Robins, 2023)

Wie zuvor bereits erwähnt, ist lediglich eine der beiden Optionen, das Factual, in der Realität beobachtbar - entweder das Treatment wird verabreicht, oder nicht. Die Formel besagt, dass Factual und Counterfactual verschieden voneinander sein müssen, um zu beweisen, dass ein kausaler Effekt vorliegt - per Definition ist dies jedoch im individuellen Fall nicht messbar. Kausale Effekte können dennoch untersucht werden, indem der durchschnittliche kausale Effekt betrachtet wird. Der durchschnittliche kausale Effekt ist der Effekt, welcher in der Grundgesamtheit zu beobachten wäre. Mathematisch lässt sich der durchschnittliche kausale Effekt wie folgt ausdrücken:

Ein durchschnittlicher kausaler Effekt liegt vor, wenn

$$E(Y^{a=1}) \neq E(Y^{a=0}) \tag{3}$$

(Hernan & Robins, 2023)

Das *E()* steht dabei für den statistischen Erwartungswert. Die Formel besagt also, dass ein durchschnittlicher kausaler Effekt vorliegt, wenn der Erwartungswert des Outcomes bei Verabreichung des Treatments sich von dem Erwartungswert des Outcomes bei Nicht-Verabreichung des Treatments unterscheidet. Interpretieren lässt sich das für das obige Beispiel als: „Wenn die durchschnittliche Erwartung einer Genesung durch die Anwendung des Treatments anders ist als ohne das Treatment, dann lässt sich von einem kausalen Effekt sprechen“. In der Realität wird allerdings selten die Grundgesamtheit untersucht, da dies häufig einen großen Aufwand darstellt oder teilweise schlichtweg nicht möglich ist. Es wird mit Stichproben gearbeitet (Chambers, 2003). Es ist jedoch zu berücksichtigen, dass die Untersuchung von Stichproben eine Fehlerquelle für Zufallsfehler darstellt. Um valide Schlussfolgerungen ziehen zu können, ist es daher wichtig, sich dieser Beschränkung bewusst zu sein und die Stichprobe repräsentativ für die Grundgesamtheit zu wählen (Chambers, 2003).

Eine weitere Quelle für Zufallsfehler sind nicht-deterministische Factuals bzw. Counterfactuals (Hernan & Robins, 2023). Nicht-deterministisch bedeutet in diesem Kontext, dass das Ergebnis eines Treatments nicht vorbestimmt ist. Liegen $Y^{a=1}$ und $Y^{a=0}$ beispielsweise eine stochastische Komponente zugrunde, hat ein Treatment nur eine Wahrscheinlichkeit zu wirken. Wird in 100 hypothetischen Fällen dieselbe Person *X* einem Treatment unterzogen, könnte das Treatment in nur 80 dieser Fälle wirksam sein - obwohl alle anderen Faktoren identisch gehalten sind. Das Treatment hat in diesem Fall also eine hohe Wahrscheinlichkeit zu wirken, aber es ist nicht garantiert. Nicht-deterministische Treatments führen demnach eine Unsicherheit in die Betrachtung ein und erschweren dadurch die isolierte Betrachtung kausaler Effekte. Es wurde

vermutet, dass Kausalzusammenhänge in der Realität nicht-deterministisch sein können (Yarlett, 2002). So lässt sich beispielsweise beim regelmäßigen Rauchen von Zigaretten kein deterministischer Zusammenhang zur Entstehung von Lungenkrebs feststellen, dennoch besteht ein Anhaltspunkt für einen kausalen Zusammenhang (Davey Smith, 2009), (Resnik & Vorhaus, 2006). Die zwei zuvor erläuterten Quellen für Zufallsfehler lassen sich zum Abschluss dieses Kapitels wie folgt festhalten:

- Das Verwenden von Stichproben bedeutet, dass nicht $E(Y^{a=1})$ und $E(Y^{a=0})$ bzw. $P(Y^{a=1})$ und $P(Y^{a=0})$ beobachtet werden können, sondern

$$\tilde{P}(Y^{a=1}) \text{ und } \tilde{P}(Y^{a=0}) \tag{4}$$

- Die Möglichkeit nicht-deterministischer Outcomes unter Bedingung von $a=0$ bzw. $a=1$ besteht.

(Hernan & Robins, 2023)

2.1.3 Confounding

In der Kausalen Inferenz geht es darum, den kausalen Effekt, den ein Treatment D potenziell auf einen Outcome Y hat, zu ermitteln. In der Realität spielen jedoch häufig viele Faktoren zusammen und es ist nicht offensichtlich, ob ein Outcome Y ausschließlich von dem Treatment D beeinflusst wurde (Hernan & Robins, 2023). Gibt es eine Variable, welche sowohl das Treatment, als auch den Outcome beeinflusst, so wird von einem Confounder gesprochen (VanderWeele & Shpitser, 2013). Wird beispielsweise in einer Beobachtungsstudie, also einer Studie in welcher die Daten nicht aus einem randomisierten Experiment stammen, der Effekt von Rauchen auf Herzkrankheiten untersucht, sind mögliche Confounder denkbar.

Im folgenden Beispiel ist Rauchen das Treatment, Übergewicht ein Confounder und die Herzkrankheit der Outcome. Angenommen Übergewicht hat einen negativen Effekt auf die Herzgesundheit, bzw. erhöht die Wahrscheinlichkeit eine Herzerkrankung zu erleiden. Beeinflusst Übergewicht jetzt zusätzlich noch das Rauchen, so liegt ein Confounder vor. Übergewicht könnte ein Indikator für die Vernachlässigung der eigenen Gesundheit sein. Menschen, die ihre Gesundheit vernachlässigen, sind möglicherweise weniger abgeschreckt von den negativen Konsequenzen des Rauchens. In diesem fiktiven Beispiel wäre die Wahrscheinlichkeit, dass übergewichtige Menschen zudem auch rauchen höher, als bei nicht übergewichtigen Leuten. In diesem Fall wirkt der Confounder „Übergewicht“ über zwei Pfade auf Y. Zum einen die direkte Auswirkung, die Übergewicht auf den Outcome hat, weil das Risiko für Herzkrankheiten dadurch erhöht wird und zum anderen die indirekte Auswirkung - da Übergewicht die Wahrscheinlichkeit erhöht zu rauchen, erhöht es zudem indirekt auch die Wahrscheinlichkeit eine Herzerkrankung zu erleiden.

Im Falle von Confounding können kausale Effekte nicht präzise bestimmt werden (Greenland et al., 1999b). Es ist unklar, wie viel des Einflusses, der auf Y wirkt, auf den kausalen Effekt des Treatments D und wie viel davon auf den Confounder X zurückzuführen ist. Die Verbindung zwischen Treatment und Outcome über den Confounder wird Backdoor-Path genannt. Bleibt ein Backdoor-Path unerkannt oder unberücksichtigt, kann dies zu einer Verzerrung des geschätzten kausalen Effekts führen (Greenland et al., 1999a). Um an dieser Stelle dennoch einen kausalen Effekt des Treatments bestimmen zu können, muss auf den Confounder X, in diesem Fall das Übergewicht, bedingt werden (Greenland et al., 1999a). „Bedingen“ (oder auch „kontrollieren“) bedeutet in der Statistik, eine Variable konstant zu halten, sodass keine Ungenauigkeit durch Änderungen in der Variable mehr entstehen kann (Durrett, 2019). Im vorigen Szenario könnten lediglich Personen, die einen BMI von Z haben, in die Betrachtung eingehen. Da der BMI und damit auch der Faktor „Übergewicht“ nun bei allen Probanden identisch ist, wurde der Confounding-Bias behoben.

2.2 Double Machine Learning

Nachdem zuvor die theoretische Grundlage im Bereich der Kausalen Inferenz gelegt wurde, soll nun das im Rahmen dieser Arbeit angewandte Double Machine Learning Framework erklärt werden. Machine Learning generell ist ein relevantes Werkzeug in Zeiten von Big Data (Wang & Alexander, 2016). Riesige Datenmengen sind heutzutage in vielen Bereichen zum Standard geworden, was bei klassischen statistischen Methoden, wie beispielsweise der linearen Regression, zu Problemen führen kann. Während die lineare Regression zwar von mehr Beobachtungen in den Daten profitiert, neigt die lineare Regression gerade bei hochdimensionalen Daten (d.h. ein hohes Verhältnis von Variablen zu Beobachtungen) zum Overfitting, also dem „Auswendiglernen“ der Datenpunkte (Hawkins, 2003). Andere Machine Learning Modelle profitieren hingegen häufig deutlich von großen Datenmengen, da sie darauf ausgelegt sind, komplexe, nicht-lineare Beziehungen zwischen Variablen zu erlernen, ohne starke Annahmen über die zugrundeliegende Verteilung der Daten zu treffen. Zudem haben viele Machine Learning Algorithmen Mechanismen, mit denen Overfitting entgegengewirkt werden kann (Zhou et al., 2017).

Datensätze, welche für die Untersuchung von kausalen Effekten verwendet werden, sind in der Regel ebenfalls umfangreich, da alle im Kausalzusammenhang stehenden Variablen in den Datensatz aufgenommen werden müssten (Hernan & Robins, 2023). Deshalb ist es naheliegend, dass auch in diesem Bereich Machine Learning einen Einzug findet. Es gibt grundlegend zwei Ansätze Machine Learning für die Schätzung kausaler Effekte anzuwenden: „Naive Machine Learning“ und „Double Machine Learning“. Naive Machine Learning bedeutet, dass ein Machine Learning Modell verwendet wird, um anhand der Kovariablen X den Effekt von dem Treatment D auf den Outcome Y zu schätzen. Dabei entsteht

potenziell der sogenannte Regularization Bias. Ein Bias ist eine systematische Abweichung bzw. Verzerrung der Ergebnisse vom wahren Sachverhalt (Georgii, 2009). Double Machine Learning hingegen ist ein Framework, das darauf abzielt, Machine Learning Algorithmen für Kausale Inferenz zu verwenden, ohne dabei einen Bias zu erzeugen (Chernozhukov et al., 2018). In den folgenden Abschnitten wird erst die Problematik von Naive Machine Learning - konkret, der Regularization Bias - erläutert. Im Anschluss daran wird Double Machine Learning aufbauend auf dem vorigen Abschnitt erklärt.

2.2.1 Naive Machine Learning und Regularization Bias

Wie bereits erwähnt, ist der Regularization Bias Grund dafür, dass Naive Machine Learning verzerrte Ergebnisse bei der Schätzung kausaler Effekte liefern kann. So kann das Auslassen von Confoundern dazu führen, dass das verwendete Modell den zugrundeliegenden Kausalzusammenhang nicht ausreichend abbildet (Hernan & Robins, 2023). Regularization ist eine gängige Methode, um die Komplexität der Machine Learning Algorithmen zu „bestrafen“ und dadurch Overfitting bzw. die Varianz des Modells zu reduzieren (Tibshirani, 1996). Die Problematik daran ist, dass dadurch ein Bias in die Ergebnisse eingeführt wird. Die Einführung eines Bias bedeutet in diesem Kontext, dass der zugrundeliegende Sachverhalt nun mit einem Modell abgebildet wird, das simpler ist, als der eigentliche Sachverhalt. Da das Modell potenziell zu simpel für den Sachverhalt ist, kann es diesen nicht akkurat approximieren - das wird auch Underfitting genannt. Sowohl Overfitting, als auch Underfitting (durch Regularization) führen zu verzerrten Modellen bei der Anwendung von Naive Machine Learning (Chernozhukov et al., 2018). Es gilt, den Kompromiss zwischen Bias und Varianz zu finden, bei dem der Generalisierungsfehler im Modell am geringsten ist - dieser Zusammenhang ist auch als Bias-Varianz-Tradeoff bekannt (Kohavi & Wolpert, 1997). Der durch Anwendung von Naive Machine Learning entstehende Regularization Bias kann auch durch Simulationsdaten visualisiert werden. Es gilt folgendes Teil-lineares Modell:

$$\begin{aligned} y_i &= \theta_0 d_i + g_0(x_i) + \zeta_i, \quad & \zeta_i &\sim \mathcal{N}(0,1), \\ d_i &= m_0(x_i) + v_i, & v_i &\sim \mathcal{N}(0,1). \end{aligned} \tag{5}$$

Es soll der kausale Koeffizient θ_0 geschätzt werden. Dabei sind die Kovariablen durch $x_i \sim \mathcal{N}(0, \Sigma)$ gegeben und $m_0(x_i)$ und $g_0(x_i)$ sind die Störfunktionen. Da es sich um Simulationsdaten handelt, können die Parameter Σ und θ_0 vom Anwender festgelegt werden.

In der Abbildung 1 wurde Naive Machine Learning auf den simulierten Datensatz angewandt. Dabei steht $\hat{\theta}$ für die Schätzung des kausalen Effekts, θ für den reellen kausalen Effekt und $\hat{\sigma}$ für die Standardabweichung. Die Gleichung $(\hat{\theta} - \theta)/\hat{\sigma}$ beschreibt also die standardisierte Abweichung des geschätzten kausalen Effekts von dem wirklichen kausalen Effekt. Da in diesem Fall der reelle kausale Effekt bekannt

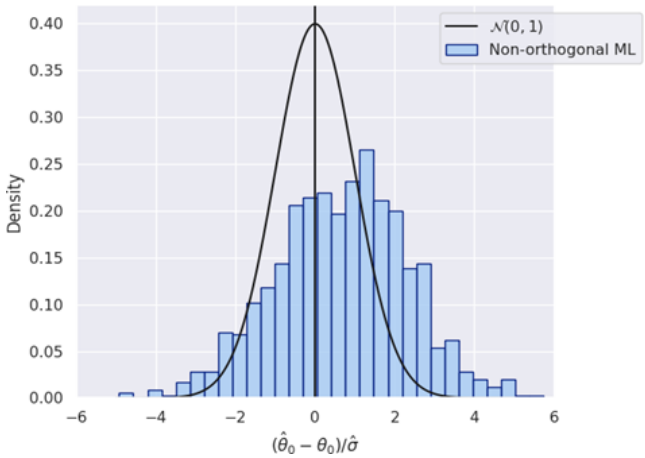

Abbildung 1: Regularization Bias durch Anwendung von Naive Machine Learning (Quelle: (Bach et al., 2022))

ist, kann so die Fähigkeit von Modellen und Frameworks evaluiert werden, diesen zu schätzen. Liegt in dem Modell kein Bias vor, wäre zu erwarten, dass das Histogramm wie der Zufallsfehler verteilt ist (in diesem Fall normalverteilt) - in der Abbildung ist jedoch deutlich eine Abweichung von dieser Normalverteilung zu sehen. Die Anwendung von Naive Machine Learning riskiert einen Bias einzuführen, welcher die resultierenden Schätzungen der kausalen Effekte verzerrt (Chernozhukov et al., 2018).

2.2.2 Double/Debiased Machine Learning

In diesem Abschnitt soll das Double Machine Learning Framework erläutert werden. Double Machine Learning (gelegentlich auch Debiased Machine Learning) ist eine mehrstufige Methode, um robuste Schätzungen für kausale Effekte zu erzeugen (Chernozhukov et al., 2016). Die folgende Gleichung beschreibt das Modell:

$$Y_i = d_i \alpha_0 + x_i' \beta + \epsilon \tag{6}$$

- Y_i ist der Outcome
- d_i ist das Treatment
- z_i sind mögliche Confounder
- x_i ist die Liste aller Kontrollvariablen (schließt z_i und Interaktionen ein)
- α ist der Koeffizient von d_i bzw. der kausale Effekt, den es zu schätzen gilt
- ϵ ist ein Störterm

Zunächst wird der Datensatz in k verschiedene Sample unterteilt - das nennt sich Crossfitting oder auch Sample Splitting und dient dazu, Overfitting zu vermeiden. Auf jedem der Sample werden zwei Machine Learning Modelle trainiert. Das erste Modell schätzt die Treatment-Variable D, das zweite Modell die Outcome-Variable Y. In beiden

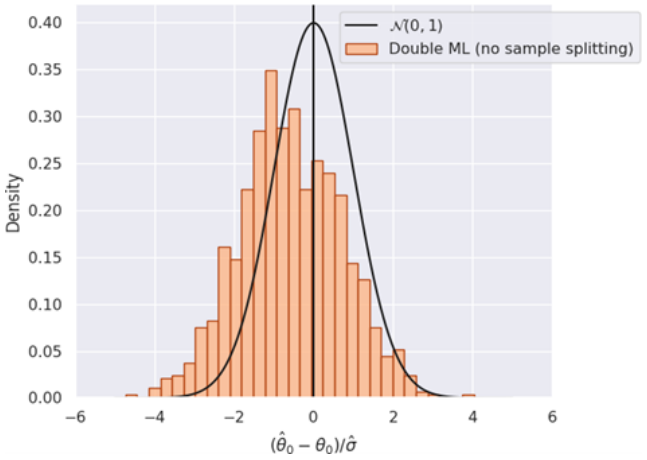

Abbildung 2: Double Machine Learning - ohne Crossfitting (Quelle: (Bach et al., 2022))

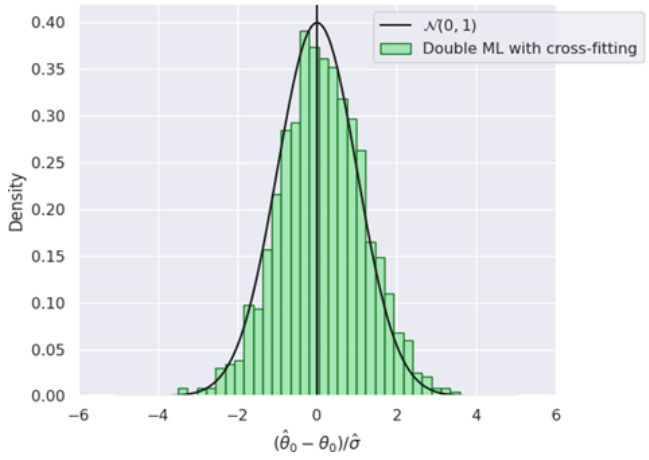

Abbildung 3: Double Machine Learning - mit Crossfitting (Quelle: (Bach et al., 2022))

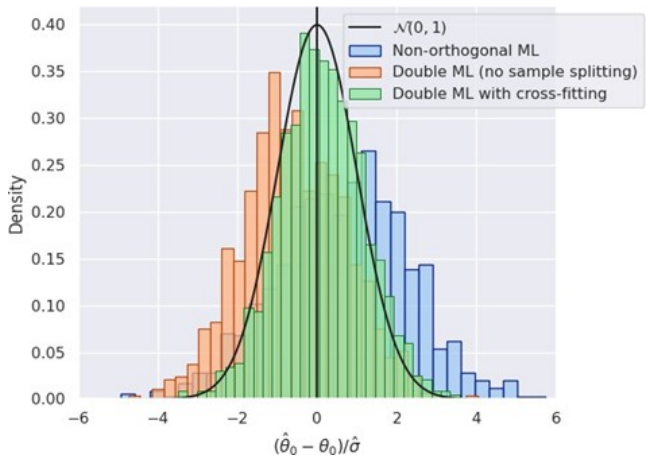

Abbildung 4: Vergleich: Naive ML, Double ML, Double ML Crossfitting (Quelle: (Bach et al., n. d.))

Fällen werden die Schätzungen auf der Grundlage der Kontrollvariablen X getroffen. Die Schätzungen werden als $\hat{D}_k$ und $\hat{Y}_k$ respektive gekennzeichnet. Anschließend werden die sogenannten Residuen $r_{D_k} = D_k - \hat{D}k$ und $r_{Y_k} = Y_k - \hat{Y}_k$, also die Differenzen zwischen den geschätzten und den eigentlichen Werten, errechnet. Nachdem die Residuen für jedes Sample errechnet wurden, wird ein weiteres Verfahren (in der Regel OLS) angewandt - dabei werden die Residuen der Treatment-Variable r_{D_k} als die unabhängige und die Outcome-Residuen r_{Y_k} als die abhängige Variable verwendet. Der von OLS geschätzte Koeffizient α_k ist der kausale Effekt, den das Treatment (im Sample) auf den Outcome hat. Die Ergebnisse aller k Samples werden zusammengeführt, um eine robustere Schätzung des kausalen Effekts α zu erhalten. Zusätzlich werden noch Hypothesentests angewandt, um die statistische Signifikanz der Ergebnisse zu bewerten. Sind die Ergebnisse signifikant, so kann von einem kausalen Effekt des Treatments auf den Outcome gesprochen werden (Chernozhukov et al., 2016, 2018).

In der Abbildung 2 ist zunächst einmal die Performance von Double Machine Learning ohne die Verwendung von Crossfitting dargestellt. Es ist erkennbar, dass die Qualität der Schätzungen sich im Vergleich zum naiven Ansatz verbessert hat - der Bias ist kleiner. Jedoch ist nach wie vor eine Verzerrung der Ergebnisse ersichtlich. Dieser Bias entsteht in diesem Fall durch Overfitting.

Um den Bias zu beheben, kann Crossfitting angewandt werden. Die Ergebnisse für Double Machine Learning unter der Verwendung von Crossfitting können in der Abbildung 3 gesehen werden.

Bei dem Vergleich der drei Verfahren in der Abbildung 4 ist ersichtlich, dass Double Machine Learning unter Verwendung von Sample Splitting die besten Ergebnisse liefert. Diese Erkenntnis ist konsistent mit den Ergebnissen von Chernozhukov et al. in ihrer Publikation „Double/debiased machine learning for treatment and structural parameters“ (Chernozhukov et al., 2018). Im Rahmen dieser Arbeit wird ebenfalls mit Double Machine Learning unter Verwendung von Crossfitting gearbeitet.

3 Oregon Health Insurance Experiment

Das Oregon Health Insurance Experiment ist eine amerikanische Studie, bei der die Auswirkungen von einer Ausweitung öffentlicher Krankenversicherung auf eine weitreichende Auswahl an Variablen, wie zum Beispiel Gesundheit, Wohlbefinden und finanzielle Lage der Teilnehmer, untersucht wurde. Die Studie wurde im Jahr 2008 in Oregon als randomisiertes Experiment mit Kontrollgruppe durchgeführt. Die erhobenen Daten und die Ergebnisse der Studie sind öffentlich frei verfügbar und wurden bereits in Studien näher untersucht. Ziel der wissenschaftlichen Arbeiten ist es, mehr über die Vor- und Nachteile der Ausweitung öffentlicher Krankenversicherung in den Vereinigten Staaten zu sammeln (Baicker & Finkelstein, 2014). In den kommenden Abschnitten wird das Experiment und die damit verbundenen Daten genauer betrachtet, um eine

zielgerichtete und nachvollziehbare Analyse der kausalen Effekte durchzuführen. Für die korrekte Anwendung Kausaler Inferenz ist es von hoher Bedeutung, den untersuchten Datensatz und seine Zusammenhänge zu verstehen. Daher werden im ersten Unterkapitel der Hintergrund und das Lotterieverfahren des OHIE beschrieben. Im zweiten Abschnitt folgt eine Erläuterung der verschiedenen Datensätze des Experiments und einer Selektion an relevanten Variablen.

3.1 Lotterie

Der goldene Standard des wissenschaftlichen Arbeitens ist ein randomisiertes Experiment (Chalmers et al., 1981). Eine nicht-zufällige Auswahl der Probanden einer Studie könnte potenziell einen Bias in die Ergebnisse einführen - beispielsweise durch den sogenannten Selection Bias (gelegentlich auch Sampling Bias). Dieser tritt auf, wenn die untersuchte Teilgruppe die Grundgesamtheit nicht akkurat widerspiegelt. Haben bestimmte Personengruppen eine höhere Wahrscheinlichkeit in das Sample einzugehen als andere, so verzerrt dies das Verhältnis in Relation zur Grundgesamtheit und verzerrt somit auch die Ergebnisse (Heckman, 1977). Das Erheben von zufälligen Stichproben gewährleistet in der Theorie ein Sample frei von Bias, weshalb Ziel in der Konzeption einer jeden Studie sein sollte, die Probanden so randomisiert wie möglich auszuwählen (Chalmers et al., 1981).

Beim Oregon Health Insurance Experiment wurden die Probanden durch ein Lotterieverfahren ausgewählt, um eine zufällige Auswahl zu gewährleisten. Die Studie umfasste etwa 90.000 Personen, die auf der Warteliste eines ansonsten geschlossenen MedicAid-Programms standen, von denen ca. 30.000 Menschen in insgesamt acht Verlosungen ausgewählt wurden. Die ausgewählten Personen hatten die Möglichkeit, sich für eine öffentliche Krankenversicherung namens Oregon Health Plan Standard zu bewerben. Oregon Health Plan Standard richtet sich dabei an einkommensschwache Erwachsene im Alter von 19-64 Jahre, welche keinen Anspruch auf andere MedicAid-Programme oder ähnliche öffentliche Krankenversicherungen in Oregon haben. Um an der Lotterie teilnehmen zu können, mussten die Kandidaten bestimmte Kriterien erfüllen, wie z.B. einen Wohnsitz in Oregon, ein Einkommen unterhalb der Armutsgrenze und Ersparnisse unter 2.000 Dollar, sowie ein andauernder Zeitraum von mindestens sechs Monaten, währenddessen der Kandidat nicht krankenversichert war. Zusätzlich zur eigenen Bewerbung bestand auch die Möglichkeit weitere Haushaltsmitglieder in die Bewerbung einzubeziehen (Baicker & Finkelstein, 2014).

Zunächst wurde die Liste der Bewerber von Duplikaten bereinigt. Anschließend wurden Bewerber, die nicht den festgelegten Kriterien entsprachen, aussortiert. Dazu gehörten Personen, die vor Ablauf der Studie das maximale Alter von 64 Jahren erreicht hätten, Personen, die nicht in Oregon wohnhaft waren und Bewerber, die vor Beginn der Studie bereits verstorben waren. Nach der Bereinigung des Datensatzes blieben noch 74.922 zulässige Bewerber übrig, unter denen Plätze im Oregon Health Plan Standard verlost wurden. Insgesamt haben 29.834 Menschen in der Lotterie gewonnen und erhielten somit die Möglichkeit, sich für das MedicAid-Programm zu bewerben. Die Möglichkeit auf Bewerbung für das MedicAid-Programm wurde zusätzlich auch auf jedes Haushaltsmitglied der Gewinner ausgeweitet. Wenn zum Beispiel eine Person aus einem Zweipersonenhaushalt in der Lotterie gezogen wurde, konnten sowohl der Gewinner als auch sein Lebenspartner oder Mitbewohner sich für MedicAid bewerben, selbst wenn Letzterer nicht an der Verlosung teilgenommen hatte (Baicker & Finkelstein, 2014). Streng genommen werden dadurch die Annahmen des randomisierten Experiments verletzt, da nicht mehr alle Teilgruppen eine identische Wahrscheinlichkeit haben, im Datensatz zu erscheinen (Heckman, 1977). Hat beispielsweise jede Person in einem vierköpfigen Haushalt separat an der Verlosung teilgenommen, so hätte es gereicht, wenn eine Person als Gewinner gezogen wird, damit der gesamte Haushalt sich für das MedicAid-Programm bewerben könnte. Demnach hätte jede Person des Haushalts eine viermal so hohe Chance, in die Versuchsgruppe einzugehen, wie eine einzeln teilnehmende Person. Unter Beachtung dieser Tatsache ist die Versuchsgruppe des Oregon Health Insurance Experiments nur dann zufällig verteilt, wenn auf die Anzahl der Mitglieder eines Haushalts, welche auf der Warteliste stehen, bedingt wird. Es ist also relevant, bei der Auswertung des Datensatzes auf die angemeldeten Haushaltsmitglieder zu bedingen, um einen Sampling Bias zu vermeiden. Da diese Variable jedoch im Datensatz erfasst wurde, ist es möglich diese bei der Auswertung der Ergebnisse einzubeziehen. Da dieser Umstand in dieser Arbeit berücksichtigt und kontrolliert wurde, ist jedoch davon auszugehen, dass kein Bias durch den Versuchsaufbau in die Ergebnisse eingegangen ist.

3.2 Überblick über die Daten

In diesem Abschnitt wird der Fokus auf die bei den Teilnehmern gemessenen Variablen gelegt, welche im Rahmen mehrerer Umfragen erhoben wurden. Wenn mehr Variablen bei der Untersuchung kausaler Effekte erfasst werden, ist es wahrscheinlicher, alle relevanten Variablen für den Kausalzusammenhang zu berücksichtigen (Hernan & Robins, 2023). Der Datensatz des OHIE ist so groß, dass im Rahmen dieses Kapitels nicht jede Variable separat erklärt werden kann. Stattdessen wird eine Übersicht über eine Auswahl an potenziell relevanten Variablen gegeben, um ein allgemeines Verständnis für die später folgenden Auswertungen zu gewährleisten. Es ist jedoch möglich, dass nicht-erläuterte Variablen Teil späterer Analysen werden. Um einen umfassenden Überblick über die Variablen des OHIE und ihre Bedeutung zu erhalten, wird empfohlen, den Userguide der Studie zu konsultieren. In diesem finden sich ebenfalls deskriptive Analysen jeder einzelnen Variable. Die Daten des OHIE sind in einem relationalen Datenbankformat in mehrere Datensätze unterteilt. Jeder Teilnehmer wurde mit einer einzigartigen Identifikationsnummer versehen, die als Primärschlüssel in allen

Datensätzen verwendet wird. Diese ist in der Variable **person_id** gespeichert.

Der Datensatz *oregonhie_descriptive_vars.dta* enthält demographische Informationen der Teilnehmer sowie Angaben darüber, ob sie in der Lotterie ausgewählt wurden und ihren Bewerbungsstatus. Wichtige Variablen aus dem Datensatz sind die **person_id**, die **household_id** und das **treatment**. Letzteres ist eine Binärvariable, die Auskunft darüber gibt, ob ein Teilnehmer in der Lotterie gewonnen hat. Drei Variablen des Datensatzes geben verschiedene Zeitpunkte, die im Zusammenhang mit der Lotterie stehen an: **dt_notify_lottery** ist das Datum, an dem Gewinner der Lotterie darüber benachrichtigt wurden, dass sie sich für das MedicAid Programm bewerben können. **dt_app_decision** dokumentiert das Datum, an dem über die Bewilligung der Bewerbung eines Antragstellers entschieden wurde. Da der Versicherungsschutz über das MedicAid-Programm rückwirkend ausgestellt wurde, beschreibt **dt_retro_coverage** das Datum, ab dem der Versicherungsschutz einer Person aktiv wurde - da die gesamte Lotterie in mehreren Etappen ausgeführt wurde, kann sich das Datum in Abhängigkeit davon, welche Iteration der Lotterie gewonnen wurde, unterscheiden. Für die Gewinner der ersten Auslosung am 10. März 2008 wurde der Versicherungsschutz beispielsweise rückwirkend für alle medizinischen Kosten aktiviert, die am oder nach dem 11. März 2008 entstanden sind, obwohl die eigentlichen Bewerbungen der Teilnehme erst zu einem späteren Zeitpunkt bewilligt wurden. Wenn ein Teilnehmer im Studienzeitraum, also nach Genehmigung seiner Bewerbung und vor dem 1. Januar 2010, verstorben ist, wird dies in der Variable **postn_death** vermerkt. Dabei werden jedoch nur Todesfälle innerhalb von Oregon erfasst. Es ist theoretisch denkbar, dass verstorbene Personen, die außerhalb von Oregon gestorben sind, nicht gekennzeichnet sind. (Baicker & Finkelstein, 2014).

Bei *oregonhie_survey0m_vars.dta* handelt es sich um den Datensatz, welcher die Ergebnisse einer ersten Umfrage beinhaltet. Die erste von drei schriftlichen Umfragen wurde zwischen Juni und November im Jahr 2008 durchgeführt. Eine Stichprobe von 58.405 Personen bekam die Umfrage zugeschickt - davon waren 29.589 Menschen Teil der Versuchsgruppe, die anderen 28.816 gehörten zur Kontrollgruppe. Insgesamt haben 26.423 Teilnehmer auf die Umfrage geantwortet. Die nächste der Umfragen wurde sechs Monate nach der initialen Befragung durchgeführt - der dazugehörige Datensatz heißt *oregonhie_survey6m_vars.dta*. Die Durchführung sowie die abgefragten Variablen sind identisch mit denen aus der ersten Umfrage. 6.359 Teilnehmer antworteten auf diese Umfrage (Baicker & Finkelstein, 2014). Die letzte der drei Umfragen, deren Ergebnisse im Datensatz *oregonhie_survey12m_vars.dta* gespeichert sind, wurde 12 Monate nach der initialen Befragung durchgeführt. Bei den Umfragen wurde ein monetärer Anreiz gesetzt, um die Teilnahmequote zu erhöhen - 5 Dollar bei Abschluss der Umfrage und zusätzlich die Chance 200 Dollar bei einer Verlosung zu gewinnen. Da die Antwortquote mit 36% recht gering war, wurde noch ein Nachbearbeitungsprotokoll gestartet - von den Personen, welche nicht geantwortet haben, wurden 30% ausgewählt und mehrmals telefonisch sowie postalisch kontaktiert und mit weiteren monetären Anreizen angesprochen. Dadurch konnte die Zahl der eingegangenen Antworten auf insgesamt 23.777 angehoben werden, was einer Antwortquote von 41% bzw. einer gewichteten Antwortquote von 50% entspricht (Baicker & Finkelstein, 2014).

Es gibt zwei weitere Datensätze, welche speziell für weitere Studien angefertigt wurden. Beide Datensätze werden im Verlauf dieser Arbeit nicht untersucht, weshalb sie nicht detaillierter vorgestellt werden. Im Rahmen dieser Arbeit wird der Datensatz *oregonhie_survey12m_vars.dta* untersucht, da dieser größer ist, als als der Datensatz *oregonhie_survey6m_vars.dta*. Der in dieser Arbeit bereitgestellte Python-Code zur Bereinigung und Aufbereitung der Daten ist jedoch universell für alle Datensätze des OHIE anwendbar. Mit geringen Anpassungen lassen sich sämtliche Untersuchungen der dritten Umfrage auch auf die anderen Datensätze übertragen.

4 Methodik

In den vorigen Kapiteln wurden sowohl die Grundlagen der Kausalen Inferenz, als auch das Oregon Health Insurance Experiment erläutert. In diesem Kapitel soll das Vorgehen zur der Untersuchung kausaler Effekte in dieser Arbeit beschrieben werden. Das vorliegende Kapitel hat den Zweck, dem Leser dabei zu helfen, die verschiedenen Entscheidungsprozesse nachzuvollziehen, die bei der Erstellung dieser Arbeit getroffen wurden. Insbesondere wird dabei berücksichtigt, dass die Wissenschaft der Kausalen Inferenz sowie das Teilgebiet des Double Machine Learnings noch sehr jung sind. Das Kapitel soll somit zukünftigen Arbeiten als Orientierungshilfe dienen. Im ersten Unterkapitel wird das technische Grundgerüst der Arbeit, insbesondere die Wahl der Programmiersprache und die verwendeten Packages, beleuchtet. Im zweiten Unterkapitel werden die verwendeten Modelle und der Versuchsaufbau näher beschrieben. Zur Bestimmung kausaler Effekte sind verschiedene Algorithmen und Verfahren anwendbar und es hängt von dem Datensatz sowie den dahinterliegenden Modellannahmen ab, welche davon die besten Ergebnisse liefern. Dementsprechend wird darauf eingegangen, welche Erwägungen und Annahmen im Rahmen dieser Arbeit getroffen und gegeneinander getestet wurden. Im dritten Unterkapitel wird das gesamte Preprocessing der Daten in Python detailliert dargestellt. Das Preprocessing ist wichtig, damit die Daten in einer Form vorliegen, die von den Machine Learning Modellen verarbeitet werden kann. Im vierten Unterkapitel wird RandomForest, der Machine Learning Algorithmus der in dieser Arbeit fürs Double Machine Learning verwendet wird, vorgestellt. Schließlich wird im fünften und

letzten Unterkapitel Hyperparameter-Tuning erklärt und aufgezeigt, in welchem Umfang dies im Rahmen dieser Arbeit verwendet wurde.

4.1 Python

In dieser Arbeit wird Python als Programmiersprache verwendet, um mögliche kausale Effekte im Oregon Health Insurance Experiment zu untersuchen. Die erfolgreiche Installation von Python und aller verwendeten Bibliotheken ist für die Durchführung und Reproduktion der Untersuchung notwendig. Python bietet eine Reihe an Vorteilen, die die Programmiersprache besonders geeignet für den Umgang mit Daten machen. Einer dieser Vorteile ist die Vielfalt an Open-Source-Modulen, auch Packages genannt, die mit minimalem Aufwand importiert werden können. Ein Modul ist in der Regel eine Sammlung von Funktionen, die zur Lösung thematisch verwandter Anwendungsfälle eingesetzt werden können. Die wichtigsten Packages, die in dieser Arbeit verwendet werden, sind *pandas*, *DoubleML* und *sklearn*. *pandas* ist ein Modul, das eine Vielzahl von Funktionen zur Datenverarbeitung bereitstellt. Es ermöglicht sowohl das Auslesen zahlreicher Dateiformate als auch die Umwandlung in das eigene DataFrame-Format von *pandas*. *pandas* wird hier verwendet, um die *.dta-Formate* aus dem OHIE zu lesen und die Weiterverarbeitung zu erleichtern. Das *DoubleML*-Package ist eine Implementierung des Double Machine Learning Frameworks von Chernozhukov et al. und ist sowohl für R als auch für Python verfügbar (Chernozhukov et al., 2016). Das Modul basiert auf dem bekannten Package *sklearn*, das häufig im Bereich Deep Learning und Machine Learning eingesetzt wird. Die Verwendung von *DoubleML* erleichtert die Untersuchung kausaler Effekte, da bereits implementierte Verfahren und Modelle direkt genutzt werden können. Andere verwendete Module oder Python-Syntax werden im Verlauf der Arbeit nur oberflächlich erklärt, um eine allgemeine Nachvollziehbarkeit zu gewährleisten. Für ein tieferes Verständnis wird auf die Dokumentationen der entsprechenden Packages verwiesen.

4.2 Modelle und Versuchsaufbau

Grundlegend gibt es in der Kausalen Inferenz vier verschiedene Modelle, welche potenziell implementiert werden können (Bach et al., n. d.):

- Partially Linear Regression Model (PLR)
- Partially Linear IV Regression Model (PLIV)
- Interactive Regression Model (IRM)
- Interactive IV Model (IIVM)

Welches davon für den vorliegenden Anwendungsfall geeignet ist, kann im Wesentlichen anhand von zwei Unterscheidungen getroffen werden. Die Unterscheidung besteht zum einen darin, ob das Treatment binär verteilt ist und zum anderen darin, ob Instrumentalvariablen im Modell sind, beziehungsweise vermutet werden. Im Falle eines nicht-binären Treatments ohne Instrumentalvariablen würde ein Partially Linear Regression Model verwendet werden. Da das Treatment in dem Oregon Health Insurance Experiment binär verteilt ist, lässt sich dieses Modell für die Anwendung auf den Datensatz ausschließen - es muss ein Interactive Regression Model oder ein Interactive IV Model verwendet werden. Welches der beiden Modelle bestenfalls angewandt werden sollte, hängt dabei davon ab, ob Instrumentalvariablen vorliegen (Chernozhukov et al., 2017). Instrumentalvariablen werden in der Wissenschaft verwendet, um in einem nicht-randomisierten Experiment kausale Effekte schätzen zu können. Hat eine Variable einen direkten Effekt auf eine andere Variable, kann sie als Instrumentalvariable verwendet werden - in der Regel bedeutet dies, das Instrument und Treatment eine hohe Korrelation aufweisen (Angrist & Imbens, 1995). Es ist jedoch wichtig, dass die Instrumentalvariable nicht mit dem Fehlerterm korreliert ist, da dies sonst zu einer Verzerrung der Ergebnisse führen würde (Nichols, 2006).

Liegen in dem Modell keine Instrumentalvariablen vor, so wird das Interactive Regression Model verwendet. Wird hingegen von Instrumentalvariablen ausgegangen, so sollte das Interactive IV Model verwendet werden. Im Fall des Oregon Health Experiments ist nicht eindeutig zu sagen, welches Modell besser geeignet ist. In der Theorie handelt es sich bei dem OHIE um ein randomisiertes Experiment, was die Verwendung von Instrumentalvariablen überflüssig macht. Auf der anderen Seite ist der Versuchsaufbau des OHIE nur dann ein randomisiertes Experiment, wenn auf die Anzahl an Mitgliedern pro Haushalt der Teilnehmer bedingt wird (Baicker & Finkelstein, 2014). Außerdem ist nicht ausgeschlossen, dass es bei der Datenerhebung anderweitig Fehlerquellen gab. Aus diesen Gründen werden im Rahmen dieser Arbeit beide Modelle angewandt und verglichen.

Der Aufbau ist dabei wie folgt:

A. Interactive Regression Model

Interactive Regression Models haben folgende Form

$$\begin{aligned} Y &= g_0(D,X) + U, \quad \mathbb{E}(U \mid X, D) &= 0, \\ D &= m_0(X) + V, \quad \mathbb{E}(V \mid X) &= 0 \end{aligned} \tag{7}$$

wobei das Treatment $D \in \{0,1\}$ binär verteilt ist (Bach et al., 2021). Die visualisierte Darstellung der Kausalbeziehung kann in Abbildung 5 gesehen werden. Im Oregon Health Insurance Experiment bezeichnet die Treatment-Variable, ob eine Person im Rahmen der Lotterie ausgewählt wurde. Wurde eine Person ausgewählt, so erhielt sie die Chance, sich auf MedicAid zu bewerben. Ob letztendlich tatsächlich eine erfolgreiche Bewerbung eingereicht wurde, ist in der Treatment-Variable jedoch nicht festgehalten. Ziel dieser Arbeit ist zu untersuchen, ob Public Healthcare einen kausalen Effekt auf ausgewählte Variablen hat. Demnach wird die Tatsache, dass ein Teilnehmer MedicAid erhalten hat, als kausales Treatment gesetzt. Dafür wird die Variable **approved_app** verwendet (genaueres im Kapitel 4.3). Die

ursprüngliche Treatment-Variable des OHIE wird verwendet, um die sogenannte „Intention to treat“ zu messen. Die Intention to treat misst den Effekt, den die Absicht das Treatment einzuführen, hat. Zusätzlich dazu wird das IRM ebenfalls verwendet, um den eigentlichen Treatment-Effekt zu schätzen.

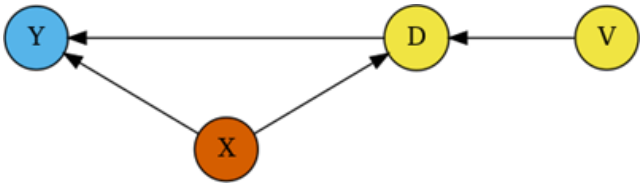

Abbildung 5. Kausal-Diagramm IRM (Quelle: (Bach et al., 2021))

B. Interactive IV Model

Interactive IV Models haben folgende Form

$$\begin{aligned} Y &= \ell_0(D,X) + \zeta, & \mathbb{E}(\zeta \mid Z,X) &= 0, \\ Z &= m_0(X) + V, & \mathbb{E}(V \mid X) &= 0 \end{aligned} \tag{8}$$

wobei das Treatment $D \in \{0,1\}$ und die Instrumentalvariablen $Z \in \{0,1\}$ binär verteilt sind (Bach et al., 2021). Die visualisierte Darstellung der Kausalbeziehung kann der Abbildung 6 entnommen werden. Mit dem IIVM wird ausschließlich der Treatment-Effekt geschätzt. Die Auswahl in der Lotterie wird dabei als Instrumentalvariable für den Erhalt von MedicAid verwendet.

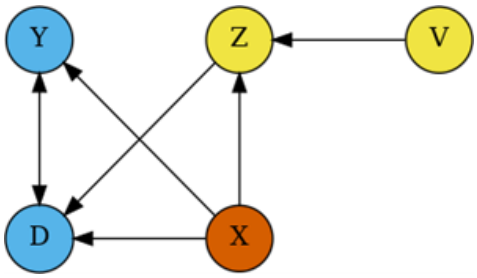

Abbildung 6. Kausal-Diagramm IIVM (Quelle: (Bach et al., 2021))

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert).

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert). Da das IIVM theoretisch mit Daten aus Beobachtungsstudien arbeiten kann, wird zusätzlich der vollständige Datensatz ohne Kontrolle auf **numhh** verwendet. Im Folgenden werden die Datensätze als „randomisierter/bedingter Datensatz“ und „vollständiger Datensatz“ betitelt.

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert). Da das IIVM theoretisch mit Daten aus Beobachtungsstudien arbeiten kann, wird zusätzlich der vollständige Datensatz ohne Kontrolle auf **numhh** verwendet. Im Folgenden werden die Datensätze als „randomisierter/bedingter Datensatz“ und „vollständiger Datensatz“ betitelt.

Für die Implementation der Modelle wird das *DoubleML*-Package verwendet. Zusätzlich zur Wahl der kausalen Modelle können Machine Learning Algorithmen spezifiziert werden. In der Praxis kann es abhängig von der Modellwahl zu leichten Abweichungen bei den Schätzungen des kau-

salen Effekts kommen. Im Rahmen dieser Arbeit wird der RandomForest-Algorithmus verwendet, weil dieser in der wissenschaftlichen Literatur vielfach erprobt ist und meist sehr robuste Ergebnisse liefert - auch ohne Hyperparameter-Tuning (Probst et al., 2019). „Robust“ bedeutet im diesem Kontext, dass der Algorithmus trotz Ausreißern, fehlenden Werten oder anderen Arten von Fehlern oder Anomalien in den Daten seine Genauigkeit beibehalten kann. Im gleichnamigen Unterkapitel werden der Algorithmus und seine Hyperparameter genauer erklärt (Syrgkanis & Zampetakis, 2020).

4.3 Datenverarbeitung

In diesem Kapitel wird der gesamte Prozess vom Import der Daten, Import der Python Libraries und die Bereinigung der Daten veranschaulicht. Alle verwendeten Packages sind in dem verwendeten Python-Environment bereits zuvor installiert werden und können deshalb mit einem *Import*-Statement abgerufen werden. Für die Replikation dieses Versuchs müsste demnach eine Python-Umgebung angelegt werden.

Zunächst werden alle benötigten Packages importiert. Wie in einem vorigen Kapitel bereits erwähnt, ermöglicht *pandas* die Handhabung und Umformung von Daten. Die *DoubleML*-Library wird für das Double Machine Learning verwendet. Das Package *sklearn* enthält den RandomForest-Algorithmus.

```
import pandas as pd
from doubleml import DoubleMLData
from doubleml import DoubleMLIIVM, DoubleMLIRM
from sklearn.ensemble import RandomForestRegressor,
    RandomForestClassifier
from sklearn.base import clone
```

Im nächsten Schritt werden die Datensätze des Oregon Health Insurance Experiments importiert und zusammengeführt. Diese liegen im Stata- bzw. *.dta*-Format vor und können mit der Funktion *pandas.read_stata()* in das *pandas.DataFrame*-Format überführt werden. Importiert wird der Datensatz mit den deskriptiven Variablen zu den Teilnehmern und die drei durchgeführten Umfragen. Anschließend werden die Datensätze zu den Umfragen jeweils mit dem Datensatz der deskriptiven Variablen zusammengeführt.

```
# import datasets as pandas.DataFrame
d_vars = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_descriptive_vars.dta")
s_0m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey0m_vars.dta")
s_6m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey6m_vars.dta")
s_12m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey12m_vars.dta")

df_0 = pd.merge(s_0m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
df_6 = pd.merge(s_6m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
df_12 = pd.merge(s_12m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
```

Anschließend gilt es, die Datensätze in eine Form zu bringen, mit der weitergearbeitet werden kann. In der Regel bedeutet dies, dass nicht-numerische Variablen in numerische Form umgewandelt werden. Zuvor sollte jedoch auf Missing Values geprüft werden. In Python werden fehlende Werte als *NaN* ("not a number") gekennzeichnet. Es gibt verschiedene Möglichkeiten mit Missing Values zu verfahren:

- Zeilen mit Missing Values löschen
- Spalten mit Missing Values löschen
- Missing Values auffüllen
- Missing Values ignorieren

Jede der Optionen hat prinzipiell verschiedene Szenarien in denen sie mehr oder weniger sinnvoll ist. Da im Oregon Health Insurance Experiment eine große Menge an *NaN* vorliegt, ist es voraussichtlich nicht sinnvoll die Werte lediglich aufzufüllen (beispielsweise mit dem Modus/Mittelwert), da dies mitunter stark die Ergebnisse verzerren könnte (Acuna & Rodriguez, 2004). Reihen bzw. Beobachtungen zu eliminieren ist ebenfalls dann sinnvoll, wenn die Quote an *NaN* hoch genug ist oder wenn die jeweilige Person bei entscheidenden Variablen keinen Wert hat - fehlt beispielsweise der Wert in der untersuchten abhängigen Variable, kann die Beobachtung nicht in die Auswertung eingehen (Acuna & Rodriguez, 2004). An dieser Stelle ist jedoch anzumerken, dass das Eliminieren von Beobachtungen immer auch mit dem Risiko verbunden ist, der Randomisierung entgegenzuwirken und dadurch Ergebnisse zu verzerren - insbesondere wenn die entfernten Beobachtungen ähnliche Charakteristiken tragen (wie z.B. viele Missing Values bzw. Missing Values an den selben Stellen) (Puma et al., 2009). Ebenfalls ist die Abwägung, ab welchem Punkt die Quote an Missing Values zu groß ist, sehr subjektiv.

Fehlende Werte können jedoch auch selbst eine Aussagekraft haben, weshalb es ebenfalls sinnvoll sein kann, diese beizubehalten. In dem Fall kann ein fehlender Wert wie eine zusätzliche Ausprägung der Variable betrachtet werden (Allison et al., 2010). Im Rahmen dieser Arbeit wird ein Mischverfahren angewandt - Beobachtungen in denen keine Werte für die abhängigen Variablen vorliegen, werden eliminiert. Missing Values an anderen Stellen werden aufgefüllt. Es ist anzumerken, dass Werte zu Variablen der Umfrage nur vorhanden sein können, wenn die entsprechende Person auch für die Umfrage ausgewählt wurde, weshalb darauf gefiltert werden muss.

Das Data Cleaning besteht aus mehreren aufeinander folgenden Schritten. Der Übersichtlichkeit halber, wurde für jeden einzelnen Teilschritt eine eigene Funktion geschrieben, welche am Ende mit *pandas.pipe()* mittels einer Pipeline auf den Datensatz angewandt werden können. Die Funktionen sind so geschrieben, dass sie universell auf jede der drei Umfragen bzw. die daraus entstandenen Datensätze angewandt werden können. Da die Variablennamen in den jeweiligen

Umfragen immer einen Suffix entsprechend der Kennung der Umfrage (0m, 6m und 12m) haben, muss der Funktion der entsprechende Suffix überreicht werden, damit Umformungen an spezifischen Spalten bzw. Variablen vorgenommen werden können. Im Folgenden werden die einzelnen Funktionen, also die Teilschritte des Data Cleanings erklärt:

```
def cache_dependant_variables(dataset:pd.DataFrame,
 name_suffix:str):
df = dataset.copy(deep=True)
df = df[df["returned" + name_suffix] == "Yes"]
cache = pd.DataFrame()
cache["health_gen" + name_suffix] = df["health_gen" +
 name_suffix].copy(deep=True)
cache["happiness" + name_suffix] = df["happiness" +
 name_suffix].copy(deep=True)
cache["med_qual" + name_suffix] = df["med_qual" +
 name_suffix].copy(deep=True)
return cache
```

Die Funktion *cache_dependant_variables* kopiert den zunächst unbereinigten Input-Datensatz und speichert die Columns der abhängigen Variablen in einem weiteren DataFrame zwischen, weil mit diesen anders verfahren wird, als mit den übrigen Variablen. Letztere werden später einem Dummy Encoding unterzogen, was mit den abhängigen Variablen nicht geschehen soll. Der Output der Funktion sind die zwischengespeicherten Variablen, welche im späteren Verlauf der Pipeline einer anderen Funktion übergeben werden.

```
def select_survey_participants(dataset:pd.DataFrame,
 name_suffix:str):
df = dataset.copy(deep=True)
df = df[df["returned" + name_suffix] == "Yes"]
return df
```

Im Schritt *select_survey_participants* wird der unbereinigte Datensatz übergeben und ausschließlich auf Zeilen begrenzt, welche Personen enthalten, welche die Umfrage beantwortet haben. Für Personen, die entweder nicht Teil des jeweiligen Samples waren, oder nicht auf die Umfrage geantwortet haben, liegen keine Werte für die Variablen der Umfrage vor, weshalb diese nicht für die Untersuchung der kausalen Effekte geeignet sind. Die Funktion gibt den neuen Datensatz, welcher nur aus Umfrageteilnehmern besteht, zurück.

```
def map_treatment_variables(df:pd.DataFrame,
 name_suffix:str):
treatment = {"Selected": 1, "Not selected": 0}
treatment2= {"Yes": 1, "No": 0}
df["treatment"] = df["treatment"].replace(treatment).
 astype(int)
df["treatment2"] = df["approved_app"].replace(
 treatment2).fillna(0).astype(int)
return df
```

Die Funktion *map_treatment_variables* wandelt die beiden Spalten mit den Treatment-Variablen vom String-Format in ein numerisches Format um. Die Spalte **treatment** steht dabei dafür, ob eine Person in der Lotterie ausgewählt wurde, während die ab hier „treatment2“ genannte Spalte angibt, wessen Bewerbung für MedicAid genehmigt wurde. Also gibt **treatment** die Intention to treat an, während **treatment2** die eigentliche Treatment-Variable im Kontext der Kausalen Inferenz widerspiegelt. Im ersten Schritt werden die Fälle, in denen ein Treatment erhalten wurde, mit einer 1 überschrieben, alle anderen Fälle mit einer 0. Bei der Variable **treatment2** gilt das auch für potenziell vorhandene Missing Values, da diese gleichbedeutend damit sind, dass die Bewerbung nicht akzeptiert wurde. Die Funktion gibt den Datensatz mit den aktualisierten Treatment-Variablen zurück.

```
def drop_problematic_columns(df:pd.DataFrame,
 name_suffix:str):
df.drop(["sample" + name_suffix, "dt_notify_lottery","
 dt_mail" + name_suffix], axis=1, inplace=True)
df.drop(["health_gen" + name_suffix, "happiness" +
 name_suffix, "med_qual" + name_suffix], axis=1,
 inplace=True)
return df
```

Durch *drop_problematic_columns* werden zum einen Spalten entfernt, die beim Survey-übergreifenden Preprocessing zu Problemen führen (primär wegen nicht uniformer Nomenklatur) und entweder bereits durch andere Variablen abgedeckt werden, oder augenscheinlich keinen Erklärungsbeitrag leisten. Im nächsten Schritt werden die abhängigen Variablen aus dem Datensatz entfernt, damit diese nicht das weitere Preprocessing unterlaufen - die Columns werden später mithilfe der bereits vorgestellten Funktion *cache_dependant_variables* wieder hinzugefügt.

```
def preprocess_object_columns(df:pd.DataFrame):
dataset_objects = df.select_dtypes(exclude=['number', '
 datetime'])
dataset_numbers = df.select_dtypes(include='number')
# split object columns in binary columns and non-binary
  columns
x = dataset_objects.apply(lambda x: x.nunique() ==2)
bool_cols = x[x==True].index
y = dataset_objects.apply(lambda x: x.nunique() !=2)
non_bool_cols = y[y==True].index
# get dummies for object columns
binary = pd.get_dummies(dataset_objects[bool_cols],
 dummy_na =True, drop_first=True).astype(int)
non_binary = pd.get_dummies(dataset_objects[
 non_bool_cols], dummy_na=True, drop_first=False).
 astype(int)
# merge the clean object and number dataframes
clean_data = pd.concat([dataset_numbers, binary,
 non_binary], axis=1)
return clean_data
```

Durch die Funktion *preprocess_object_columns* wird eine Trennung in numerische und kategoriale Variablen vorgenommen. Die kategorialen Variablen können gerade in Form von „Ja/Nein“-Fragen häufig in binärer Form vorliegen. *pandas* bietet mit der Funktion *select_dtypes()* die Möglichkeit, alle Spalten eines oder mehrerer ausgewählter Datentypen zu filtern. Da der Aufbau der Umfragen in Bezug auf die abgefragten Variablen identisch ist, sind die Schritte zur Umformung der Variablen ebenfalls (bis auf die Nomenklatur einzelner Variablen) identisch. Die numerischen Variablen können unverändert in das Modell eingehen - die kategorialen Variablen werden weiterverarbeitet. Zunächst wird in binäre und nicht-binäre Variablen getrennt. Theoretisch ist dieser Schritt nicht zwingend notwendig, bietet aber die Option mit binären Variablen anders zu verfahren. Im nächsten Schritt werden Dummy-Variablen erzeugt. Das bedeutet, dass die verschiedenen Ausprägungen der Variablen

auf mehrere Spalten aufgeteilt werden. In dem einfachen Fall, dass eine Spalte *X* die Ausprägungen „Ja“ und „Nein“ enthält, würden zwei neue Spalten angelegt werden: *X_Ja* und *X_Nein*. Für alle Zeilen, in denen in der ursprünglichen Spalte der Wert „Ja“ vorlag, wird in der neuen Spalte *X_Ja* eine 1 und in der Spalte *X_Nein* eine 0 eingetragen und vice versa. Dieses Verfahren ist skalierbar auf eine beliebige Anzahl an Antwortmöglichkeiten *k* - der limitierende Faktor ist dabei Rechenleistung und Modellperformance, welche unter zu viel Dummy-Encoding leiden kann (Fradkin & Madigan, 2003).

Um mit Dummy-Variablen alle Antwortmöglichkeiten abzubilden, werden *k-1* Spalten benötigt (Suits, 1957). Im Falle von binär verteilten Variablen, wie auch dem vorigen Beispiel, ist eine Spalte also ausreichend, um alle Antwortmöglichkeiten abzudecken. Wird nur die Spalte *X_Ja* betrachtet, bedeutet eine 0, dass in der ursprünglichen Spalte *X* ein „Nein“ stand. Da ein Großteil der Variablen jedoch auch noch *NaN* enthält und diese beibehalten werden sollen, gibt es faktisch drei Ausprägungen bei den „binären“ Variablen des Datensatzes. Mithilfe von *pandas*' Funktion *get_dummies()* werden. Die Funktion *preprocess_object_colums* erstellt Dummy-Variablen für alle übergebenen Spalten. Für die Variable **binary** werden alle binären Variablen übergeben, für die Variable **non-binary** die restlichen. Mit *dummy_na=True* wird der Funktion *get_dummies()* übergeben, dass *NaN*-Werte ebenfalls codiert werden sollen. Standardmäßig wird für jede Ausprägung einer Variable eine Dummy-Spalte erzeugt - da theoretisch jedoch nur *k-1* Spalten notwendig sind, kann mit *drop_first=True* die erste Spalte gelöscht werden. In dieser Arbeit wurde sich dazu entscheiden, *drop_first=True* für Binärvariablen zu aktivieren, um den bereinigten Datensatz bezüglich der Dimensionalität geringer zu halten. Für nicht-binäre Variablen wurde *drop_first=False* gesetzt.

Zuletzt müssen die nun allesamt numerischen Datensatz-Bestandteile wieder zusammengefügt werden. Die *pandas*-Funktion *concat()* ermöglicht es, mehrere Datensätze zusammenzufügen. Über die Achse wird angegeben, ob die Datensätze untereinander oder nebeneinander verbunden werden sollen - *axis=1* bedeutet, dass die Datensätze nebeneinander gesetzt werden. Häufig wird zum Zusammensetzen von DataFrames eine gemeinsame Variable als Schlüssel benötigt. In diesem Fall sind die Datensätze jedoch alle in der identischen Reihenfolge, in der sie sich auch befunden haben, bevor sie getrennt wurden. Deshalb kann in diesem Fall ohne einen spezifizierten Schlüssel verbunden werden. *pandas* verwendet, falls nicht weiter spezifiziert, automatisch den Index. Der Output der Funktion ist ein Datensatz, welcher im klassischen Sinne bereinigt ist und lediglich Spalten im numerischen Datentyp enthält.

```
def fill_missing_values(clean_data:pd.DataFrame):
for column in clean_data.columns:
    if clean_data[column].isna().values.any():
        clean_data[column + "wasNaN"] = clean_data[
 column].isna() * 1
clean_data.fillna(0, inplace=True)
return clean_data
```

Da der Datensatz nach dem vorigen Schritt noch Missing Values enthält, muss abgewägt werden, wie damit verfahren werden soll. Wie bereits zuvor in der Arbeit erklärt, werden die Missing Values in diesem Fall aufgefüllt. Die fehlenden Werte lagen fast ausschließlich in Spalten vor, in denen eine numerische Angabe von den Teilnehmern gefragt war - andere Spalten derselben Reihen waren sowohl vor, als auch nach den fehlenden Werten trotzdem gefüllt. Es gibt verschiedene Gründe, warum Teilnehmer Fragen in einer Umfrage unbeantwortet lassen. Beispielsweise, weil sie selbst nicht genau wissen, welche die akkurate Antwort ist oder weil die Frage für sie zu privat ist. Ebenfalls können Fragen unbeantwortet bleiben, wenn Teilnehmer nach einigen Fragen die Umfrage abbrechen. Die Variablen bzw. Fragen, welche im OHIE eine vergleichsweise hohe Quote an Missing Values aufweisen, sind vermehrt Fragen nach der Häufigkeit. Diese Fragen sind für Versuchsteilnehmer im Kontext von MedicAid thematisch zu erwarten gewesen, weshalb hier nicht davon ausgegangen wird, dass die Fragen aufgrund von Bedenken hinsichtlich der Privatsphäre nicht beantwortet wurden. Gerade bei Fragen nach der Häufigkeit könnte ein Nicht-Beantworten der Antwort „0“ gleichgesetzt worden sein. Deshalb werden durch die Funktion *fill_missing_values* sämtliche verbleibende Missing Values im Datensatz mit 0 aufgefüllt. Da jedoch durchaus die Möglichkeit besteht, dass die fehlenden Werte mehr Bedeutung hatten, werden zuvor sämtliche Spalten, welche Missing Values enthalten um eine weitere Spalte bzw. Variable erweitert. Diese zusätzliche Spalte kennzeichnet, wo zuvor in dem ursprünglichen Column Missing Values vorhanden waren, damit der potenzielle Erklärungsgehalt der fehlenden Werte erhalten bleibt.

Theoretisch wird durch das Auffüllen der Missing Values ein Bias eingeführt. Da jedoch die fehlenden Werte im Datensatz weiterhin gekennzeichnet bleiben, ist die Idee dieser Arbeit, dass der durch das Auffüllen entstandene Bias geringer ist als der, der entstünde, wenn sämtliche Reihen mit Missing Values eliminiert und der Datensatz um ein Vielfaches verkleinert werden würde.

```
def add_and_transform_dependant_columns(clean_data:pd.
 DataFrame, cache:pd.DataFrame, name_suffix:str):
health = {"excellent": 4, "very good": 3, "good": 2, "
 fair": 1, "poor": 0}
happiness = {"very": 2, "pretty happy": 1, "not too
 happy": 0}
quality = {"excellent": 4, "very good": 3, "good": 2, "
 fair": 1, "poor": 0, "no care": 5}
clean_data["med_qual_" + name_suffix + "nocare"] = (
 cache["med_qual" + name_suffix] == "no care") * 1
clean_data["health_gen" + name_suffix] = cache["
 health_gen" + name_suffix].copy(deep=True).map(health)
clean_data["happiness" + name_suffix] = cache["
 happiness" + name_suffix].copy(deep=True).map(
 happiness)
clean_data["med_qual" + name_suffix] = cache["med_qual"
  + name_suffix].copy(deep=True).map(quality)
med_qual_mean = clean_data[clean_data["med_qual" +
 name_suffix] != 5]["med_qual" + name_suffix].dropna().
 astype(int).mean()
```

```
clean_data["med_qual" + name_suffix] = clean_data["
 med_qual" + name_suffix].replace(5, med_qual_mean)
return clean_data
```

Im Anschluss an das Auffüllen der Missing Values werden durch die Funktion *add_and_transform_dependant_columns* die zuvor entfernten abhängigen Variablen wieder mit dem Datensatz verbunden. Dafür werden der Funktion der bereinigte Datensatz und die Cache-Funktion übergeben. Die Columns werden ebenfalls mit der *map()*-Funktion von *pandas* in die numerische Form überführt. Eine Besonderheit liegt beim Column **med_qual** vor. Bei diesem wird die Qualität der in Anspruch genommen medizinischen Leistungen beurteilt. Jedoch besteht auch die Möglichkeit, dass ein Teilnehmer keine Leistungen in Anspruch genommen hat. Dieser Fall wurde im ursprünglichen Datensatz mit „no care“ gekennzeichnet. In der Funktion wird der Wert „no care“ zunächst mit einem Platzhalter überschrieben und anschließend durch den Mittelwert aller Werte, die nicht „no care“ waren, ersetzt. Hintergrund ist, dass dadurch ein großer Anteil an Zeilen im Datensatz erhalten werden kann. Das Auffüllen mit dem Mittelwert verzerrt zwar die Ergebnisse, aber weniger, als wenn ein Wert von größer als 4 beziehungsweise kleiner als 0 eingesetzt werden würde. Zudem ist zu beachten, dass der Wert „no care“ lediglich überschrieben wird, wenn **med_qual** als unabhängige Variable verwendet wird. In dem Fall, dass **med_qual** als unabhängige Variable betrachtet wird, werden die Zeilen mit dem Mittelwert entfernt. Der Output der Funktion ist der bereinigte Datensatz, welcher um die nun numerischen abhängigen Variablen erweitert wurde.

```
def drop_remaining_missing_values(clean_data:pd.
 DataFrame):
clean_data.dropna(axis=0, inplace=True)
return clean_data
```

Da die in der vorherigen Formel wieder hinzugefügten Spalten das Preprocessing übersprungen haben, enthalten diese noch Missing Values, welche durch die Funktion *drop_remaining_missing_values* entfernt werden.

```
def transform_remaining_columns_to_int(clean_data:pd.
 DataFrame, name_suffix:str):
clean_data["health_gen" + name_suffix]  = clean_data["
 health_gen" + name_suffix].astype(int)
clean_data["happiness" + name_suffix] = clean_data["
 happiness" + name_suffix].astype(int)
clean_data["med_qual" + name_suffix]  = clean_data["
 med_qual" + name_suffix].astype(int)
return clean_data
```

Zuletzt müssen die unabhängigen Variablen noch mit der Funktion *transform_remaining_columns_to_int* in das Integer-Format überführt werden, da diese nach dem Mapping in *pandas* immer noch als Object-Column interpretiert werden.

```
clean_data_12m_filled =(
df_12.pipe(select_survey_participants, "_12m")
.pipe(map_treatment_variables, "_12m")
.pipe(drop_problematic_columns, "_12m")
.pipe(preprocess_object_columns)
.pipe(fill_missing_values)
.pipe(add_and_transform_dependant_columns,
 cache_dependant_variables(df_12, "_12m"), "_12m")
.pipe(drop_remaining_missing_values)
.pipe(transform_remaining_columns_to_int, "_12m")
)
```

Nun kann ein vollständig bereinigter Datensatz erzeugt werden, indem der ursprüngliche Datensatz mit der Funktion *pipe()* schrittweise den zuvor vorgestellten Funktionen übergeben wird. In jedem Schritt ist erkennbar, welcher Teil des Preprocessings angewandt wird und welche zusätzlichen Inputs, wie beispielsweise der Names-Suffix, übergeben werden. Der Prozess wurde in Abbildung 7 visualisiert dargestellt.

Da in dem weiteren Verlauf der Arbeit zwei unterschiedliche Datensätze verwendet werden, muss zu dem bereinigten Datensatz in voller Größe noch ein Datensatz erzeugt werden, bei dem auf die Variable **numhh** bedingt wird, um ein vollständig randomisiertes Experiment sicherzustellen.

```
clean_data_12m_cond = clean_data_12m_filled[
 clean_data_12m_filled["numhh_list_signed self up"] ==
 1]
```

Die Ausprägung „signed self up“ ist die am häufigsten vorkommende, weshalb auf diese Ausprägung bedingt wird, um den größtmöglichen Datensatz beizubehalten. Um Double Machine Learning mithilfe der *DoubleML*-Bibliothek auf den Datensatz anzuwenden, muss dieser in dem *DoubleMLData*-Format vorliegen. Dafür wird die Funktion *DoubleMLData()* verwendet. Als erstes wird der *pandas*-DataFrame übergeben, anschließend werden die Spalten bzw. Variablen spezifiziert. Es muss mindestens ein Treatment übergeben werden und es besteht die Option Instrumentalvariablen zu übergeben. Exemplarisch sieht dies wie folgt aus:

```
obj_dml_data = DoubleMLData(dataframe, y_col="outcome",
  d_cols="treatment", z_cols="IV_variables", x_cols="
 covariates")
```

Die resultierende Python-Variable **obj_dml_data** liegt in einer Form vor, mit der Double Machine Learning betrieben werden kann.

4.4 RandomForest

Wie im Unterkapitel zu den kausalen Modellen bereits erwähnt, wird in dieser Arbeit RandomForest als Machine Learning Algorithmus verwendet. RandomForest ist ein sogenannter Ensemble-Learner - also eine Kombination mehrerer Algorithmen, die eine höhere Leistung erzielt als die jeweils einzelnen Methoden (Breiman, 2001). In der Regel wird versucht Modelle zu kombinieren, die sich gegenseitig ergänzen bzw. die Schwächen der anderen Modelle ausgleichen. Im Falle von RandomForest werden zwei Verfahren miteinander kombiniert: Decision Trees und Bagging.

Decision Trees sind aufgrund ihrer intuitiven Interpretierbarkeit weit verbreitet. Decision Trees splitten den Datensatz anhand der Variablen so auf, dass die Schätzungen - gemessen an einer zuvor ausgewählten Metrik - verbessert werden. Dieser Vorgang nennt sich Recursive Partitioning. Um

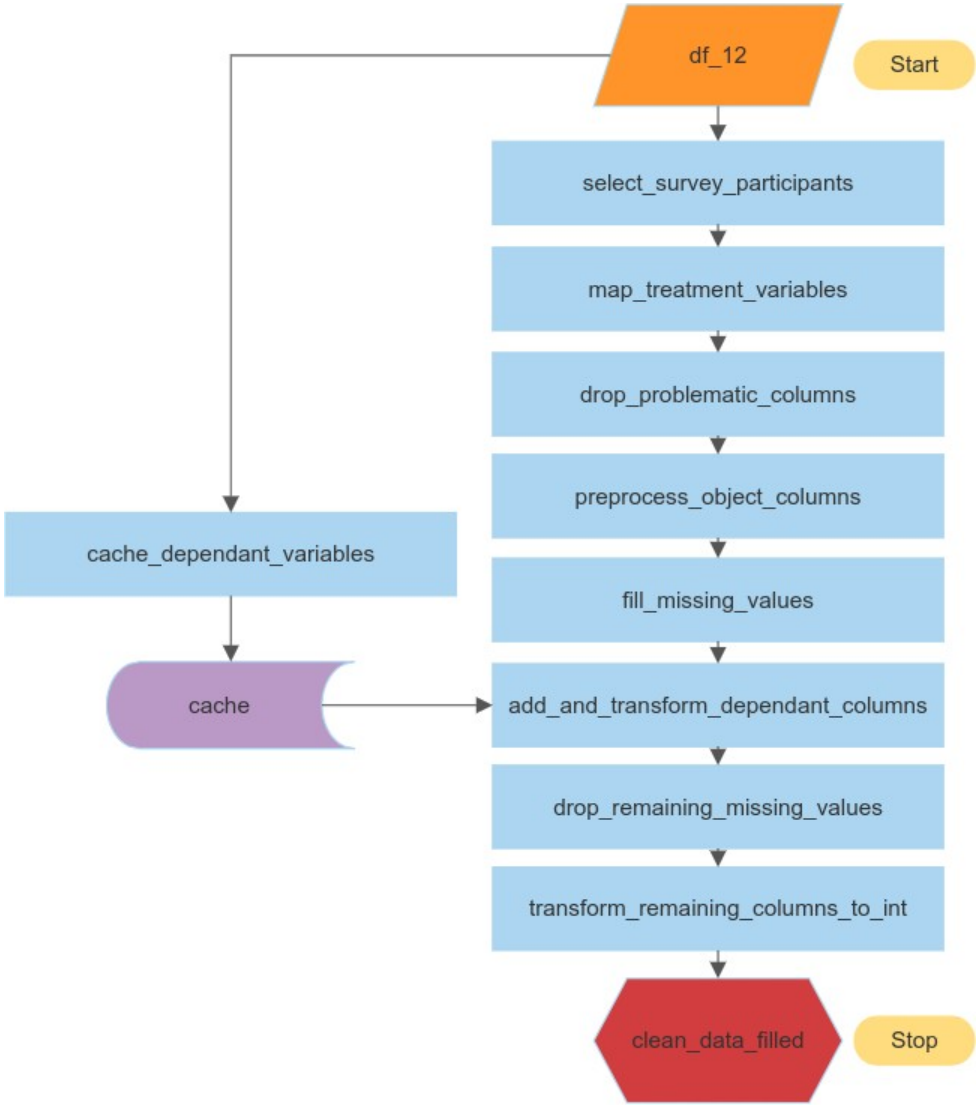

Abbildung 7: Preprocessing Flowchart (Quelle: E. D.)

einen Decision Tree zu errichten, wird Recursive Partitioning über mehrere Iterationen hinweg wiederholt - in jeder Iteration wird der Split gewählt, welcher den Schätzungsfehler am meisten reduziert (von Winterfeld & Edwards, 1986). Dieser Prozess kann theoretisch wiederholt werden, bis jeder Datenpunkt einen eigenen Knoten im Decision Tree bildet. Der Vorteil von Decision Trees ist ersichtlich: Ihre einfache Erstellung und ihre leichte Interpretierbarkeit. Der Nachteil des Modells ist jedoch, dass mit tieferen Bäumen die Wahrscheinlichkeit zunimmt, die Daten zu overfitten (Bramer, 2007). Im Extremfall, wenn jeder einzigartige Datenpunkt einen eigenen Knoten im Decision Tree hat, werden die Daten vollständig auswendig gelernt, und die Varianz des Modells ist hoch - dann wäre zu erwarten, dass die Performance auf Testdaten nicht zufriedenstellend ist.

Bagging - ausgeschrieben: Bootstrap Aggregating - ist ein Verfahren, das die Performance von Machine-Learning-Algorithmen verbessert und gleichzeitig die Varianz des Modells reduziert, indem es Overfitting verhindert. Die zentrale Idee von Bagging besteht darin, mehrere Sample aus dem Trainingsdatensatz zu ziehen, auf denen anschließend Modelle trainiert werden. Diese Modelle werden dann zu einem neuen, robusteren Modell aggregiert (Breiman, 1996). Die Besonderheit beim Ziehen der Sample ist, dass dies „mit Zurücklegen“ geschieht. Ein besonderes Merkmal des Verfahrens ist, dass die Stichproben mit Zurücklegen gezogen werden. Das bedeutet, dass jede Beobachtung potenziell in mehreren Stichproben vorkommen kann. Es ist möglich, dass einige Beobachtungen mehrfach gezogen werden, während andere gar nicht gezogen werden.

Der RandomForest-Algorithmus ist eine Kombination dieser beiden Methoden - aus dem Trainingsdatensatz werden n Sample gezogen, auf die jeweils ein Decision Tree trainiert wird. Anschließend werden diese Decision Trees zu einem Modell aggregiert. Das finale Modell ist robuster als einzelne Decision Trees (Syrgkanis & Zampetakis, 2020). Ran-

domForest hat zudem viele Hyperparameter, die optimiert werden können. Im Folgenden werden die Hyperparameter vorgestellt, welche im weiteren Verlauf der Arbeit mittels Hyperparameter-Tuning angepasst werden.

Der Hyperparameter ***n_estimators*** gibt an, wie viele Entscheidungsbäume in den RandomForest einbezogen werden und gehört damit zu den wichtigsten Parametern des Verfahrens. Grundsätzlich gilt, dass mit zunehmender Anzahl an Entscheidungsbäumen die Präzision der Ergebnisse steigt, bzw. diese robuster werden - der Grenzertrag nimmt mit steigender Zahl an Decision Trees ab. Mit ***max_features*** wird festgelegt, wie viele Variablen maximal pro Split an einem Knoten betrachtet werden, während ***max_depth*** die maximale Tiefe beziehungsweise Knotenzahl eines Entscheidungsbaumes angibt. Die maximale Anzahl an Features, die an einem Knoten betrachtet wird, erhöht die Variabilität der einzelnen Entscheidungsbäume zueinander. Wenn ***max_features*** kleiner ist als die Anzahl der Variablen im Datensatz n, wird nur ein zufälliges Subset an Variablen für jeden Split berücksichtigt. Ein geringeres ***max_features*** als n kann daher dazu beitragen, dass die Schätzungen des Random Forest robuster werden. Mit ***max_depth*** hingegen wird Overfitting reduziert, indem die Entscheidungsbäume auf eine maximale Tiefe begrenzt werden, über die sie nicht hinauswachsen können. Ist der Parameter gering genug gesetzt, wird verhindert, dass die Decision Trees jede einzelne Beobachtung auswendig lernen. Der Hyperparameter ***min_samples_leaf*** gibt an, wie viele Beobachtungen mindestens in einem Endpunkt eines Entscheidungsbaumes vorhanden sein müssen. Je nachdem, wie hoch der Wert gesetzt wird, verhindert auch dies Overfitting, da beispielsweise nicht für jede Beobachtung ein einzelner Endpunkt erzeugt werden kann. Obwohl beide Hyperparameter dazu dienen Overfitting zu reduzieren, ist es sinnvoll beide zu optimieren, da die Wirkungsweise sich unterscheidet. Zum Beispiel kann es vorkommen, dass ein Modell auf einem großen Datensatz bereits lange overfittet, aber ***min_samples_leaf*** noch nicht auslöst, da immer noch genügend Beobachtungen in jedem Endknoten enthalten sind. Der Parameter ***max_depth*** hätte möglicherweise den Entscheidungsbaum bereits abgebrochen und vice versa (Cournapeau, n. d.).

4.5 Hyperparameter-Tuning

Im Bereich des Machine Learnings gibt es bei der Arbeit mit Algorithmen häufig verschiedene Variablen, die innerhalb eines Algorithmus angepasst werden können. Diese Variablen, auch Hyperparameter genannt, beeinflussen den Lernprozess des Algorithmus und somit auch dessen Qualität (Schratz et al., 2019). Das sogenannte Hyperparameter-Tuning (manchmal auch Hyperparameter-Optimization) beschreibt den Prozess diese Parameter so auszuwählen, dass die Performance des Modells auf den individuellen Datensatz optimiert wird. Es gibt keinen allgemeingültigen „besten“ Wert für Hyperparameter. Es gibt jedoch einen theoretisch fundierten Default-Wert, der standardmäßig als Ausgangspunkt und somit als Baseline verwendet wird. Dieser Wert liefert in vielen Fällen auch ohne Hyperparameter-Tuning bereits zufriedenstellende Ergebnisse und rechtfertigt somit seine Verwendung als Default (Wong et al., 2019). Hyperparameter-Tuning ist ein Prozess, bei dem speziell für den Anwendungsfall optimale Werte für Hyperparameter bestimmt werden. Im Rahmen dieser Arbeit soll Hyperparameter-Tuning auf den zuvor erklärten RandomForest-Algorithmus angewendet werden, um so präzise Ergebnisse wie möglich zu erhalten. Im folgenden Unterkapitel werden zwei Verfahren erläutert, welche für das Hyperparameter-Tuning relevant sind.

4.5.1 Crossvalidation und Gridsearch

Nachdem zuvor bereits einige der optimierbaren Hyperparameter vorgestellt wurden, sollen im folgenden Abschnitt die Algorithmen beschrieben werden, die zur Optimierung der Hyperparameter verwendet werden. Ein gängiger Ansatz ist die Crossvalidation, bei der der Datensatz in ein Trainings- und Testset aufgeteilt wird. Eine Möglichkeit dies umzusetzen, ist die *k-fold Crossvalidation*. Dabei wird der Datensatz in k Subsets unterteilt - k kann dabei vom Anwender festgelegt werden. In jeder Iteration werden *k-1* Subsets als Trainingsdatensatz und das verbleibende Subset als Testset verwendet (Bro et al., 2008). Dabei werden in jeder Iteration verschiedene Hyperparameter-Einstellungen verwendet. Nach k Durchläufen werden die Modelle hinsichtlich ihrer Performance miteinander verglichen, um die besten Hyperparameter-Einstellungen zu ermitteln (Probst et al., 2019).

Wenn eine große Anzahl von Hyperparametern optimiert werden soll, die wiederum eine Vielzahl von möglichen Ausprägungen haben können, kann die Anzahl der zu testenden Modelle einen immensen Rechenaufwand darstellen. Hat ein Algorithmus beispielsweise lediglich einen Hyperparameter, welcher binär verteilt ist, gäbe es zwei Modelle welche gegeneinander getestet werden können. Bei fünf Parametern, welche jeweils im Bereich von 0 und 10 verteilt sein könnten, ist die Anzahl verschiedener Modelle bereits 161.051 Stück. Jedes dieser Modelle zu trainieren und zu validieren ist ein großer Rechenaufwand. Darüber hinaus ist es oft gar nicht sinnvoll, jedes denkbare Modell zu testen, da kleine Unterschiede in den Hyperparametern möglicherweise keinen signifikanten Einfluss auf die Modellleistung haben. In der Praxis ist es oftmals nicht relevant, ob 300 oder 301 Decision Trees in einem RandomForest trainiert werden (Probst et al., 2019). In diesem Fall kann das *Gridsearch*-Verfahren verwendet werden, um bestimmte Werte für die Hyperparameter auszuwählen, die gegeneinander getestet werden sollen. Werden für mehrere Parameter Werte angegeben, so werden die Kombinationen miteinander getestet und es ergibt sich ein Netz - daher auch der Name (Hsu et al., 2003). Die Werte der einzelnen Hyperparameter für *Gridsearch* können entweder manuell ausgewählt werden oder es kann ein Intervall gesetzt werden, aus welchem dann zufällig Werte gezogen werden - in diesem Fall wird von *Randomsearch* gesprochen

(Bergstra & Bengio, 2012).

4.5.2 Implementation

Im Folgenden wird die Umsetzung der beschriebenen Verfahren zum Hyperparameter-Tuning von RandomForest in Python dargestellt. Das *DoubleML*-Package hat eine eingebaute Funktion *tune()*, welche zum Tunen der Hyperparameter verwendet werden kann. Es besteht die Möglichkeit zwischen Gridsearch und Randomsearch zu wählen. Der Funktion wird ein Parameter-Grid mit allen zu testenden Ausprägungen der Hyperparameter übergeben. Nachfolgend wird die Anwendung der *tune()*-Funktion exemplarisch dargestellt.

```
ml_g = RandomForestRegressor()
ml_m = RandomForestClassifier()

param_grid = {
'n_estimators': np.array([150, 300]),
'max_depth': np.array([25, None]),
'max_features': np.array(["sqrt", None])
}

par_grids = {'ml_g': param_grid,
             'ml_m': param_grid}

obj_dml_irm.tune(par_grids, search_mode='grid_search')
fit = obj_dml_irm.fit()
```

Im obigen Code-Block werden zunächst die verwendeten Machine Learning Algorithmen definiert. Anschließend wird das Parameter-Grid definiert, welches die verschiedenen Ausprägungen der Hyperparameter enthält, die gegeneinander getestet werden sollen. Anschließend wird das Parameter Grid jedem der Machine Learner übergeben. Nachdem die Parameter fürs Tuning gesetzt wurden, muss das Grid lediglich innerhalb der *tune()*-Funktion abgerufen werden. Innerhalb der Funktion wird der Search-Algorithmus festgelegt - in diesem Fall *Gridsearch*. Abschließend wird die *fit()*-Funktion aufgerufen und der *Gridsearch*-Algorithmus wird automatisch während des Fitting-Prozesses des Modells ausgeführt.

Für diese Arbeit wurden die Parameter ***n_estimators***, ***max_depth*** und ***max_features*** als die wichtigsten Parameter ausgewählt. Da die benötigte Rechenzeit für das Hyperparameter-Tuning stark mit der Anzahl der Parameter wächst, ist schnell ein Limit erreicht, insbesondere bei großen Datensätzen. Um die Rechenanforderungen zu senken, wurde bei den IRM ein kleineres Parameter-Grid verwendet, als bei den IIVM. In der Tabelle 1 ist eine Auflistung der Werte zu sehen, welche den Modellen für die jeweiligen Hyperparameter übergeben wurden.

Alle Ausprägungen des Parameter-Grids für Interactive Regression Models finden sich auch im Grid für die Interactive IV Models wieder - letztere haben pro Hyperparameter jedoch noch jeweils eine weitere Ausprägung.

Hyperparameter	Werte für IRM	Werte für IIVM
n_estimators	300	300, 350
max_depth	25, None	10, 25, None
max_features	sqrt, None	log2, sqrt, None

Tabelle 1: Vergleich der Parameter-Grids fürs Hyperparameter-Tuning (Quelle: E. D.)

4.6 Untersuchung kausaler Effekte

Die Untersuchung der kausalen Effekte im nachfolgenden Kapitel ist von wesentlicher Bedeutung für die Beantwortung der Leitfrage dieser Arbeit, welche Effekte Public Healthcare auf die abgedeckte Bevölkerung eines Landes hat und wie stark diese sind. Nachdem in den vorangegangenen Kapiteln die theoretischen Grundlagen gelegt wurden, erfolgt in dem fünften Kapitel die Untersuchung der Auswirkungen des Treatments auf verschiedene Outcome-Variablen. Davor wird in diesem Abschnitt der dafür verwendete Code dargestellt.

Das Vorgehen für jede untersuchte Variable ist identisch: Zunächst wird der kausale Effekt der Intention to treat untersucht, d.h. wie stark der Einfluss der alleinigen Absicht ist, ein MedicAid-Programm durchzuführen. Dies dient als Baseline für die weiteren Schätzungen - sollten der Treatment-Effekt sich nicht von der Intention to treat unterscheiden, scheint es keinen signifikanten Treatment-Effekt zu geben. Dafür wird ein Interactive Regression Model verwendet. Zum Vergleich wird dasselbe Modell verwendet, um den eigentlichen Treatment-Effekt, d.h. die Aufnahme in das MedicAid-Programm, zu untersuchen. Nach Abschluss dieser beiden Untersuchungen wird der Treatment-Effekt erneut unter Verwendung des Interactive-IV-Regression-Modells geschätzt, wobei die Auswahl in der Lotterie als Instrumentalvariable verwendet wird. Dieser Ablauf wird pro Variable zweimal durchgeführt - auf jedem der Datensätze einmal. Es werden also für jede Variable, die Ergebnisse für die folgenden Fälle ermittelt:

- Datensatz: Missing Values aufgefüllt - (Vollständig)
  - Intention to treat - IRM
  - Treatment Effekt - IRM
  - Treatment Effekt - IIVM
- Datensatz: Missing Values aufgefüllt - (Bedingt auf *numhh*)
  - Intention to treat - IRM
  - Treatment Effekt - IRM
  - Treatment Effekt - IIVM

Nachfolgend ist exemplarisch der verwendete Code aufgeführt. Für jeden geschätzten kausalen Effekt wurde diese Struktur verwendet - Unterschiede gibt es lediglich bei der betrachteten Outcome-Variable, dem verwendeten Modell und der Treatment-Variable.

```
obj_dml_data = DoubleMLData(data=DataFrame, y_col=
 outcome_variable, d_cols=treatment_variable, z_cols=
 instrumental_variable)
obj_dml_iivm = DoubleMLIIVM(obj_dml_data=obj_dml_data,
 ml_g=ml_g, ml_m=ml_m, ml_r=ml_r, apply_cross_fitting=
 True, n_folds=2)

param_grid = {
    'n_estimators': np.array([300, 350]),
    'max_depth': np.array([10, 25, None]),
    'max_features': np.array(["log2", "sqrt", None])
}
par_grids = {'ml_g': param_grid,
             'ml_m': param_grid,
              'ml_r': param_grid}

obj_dml_iivm.tune(par_grids)
fit = obj_dml_iivm.fit()
print(fit.summary)
```

In der ersten Zeile wird das *DoubleMLData*-Objekt mithilfe der Funktion *DoubleMLData()* erzeugt. Die Funktion hat mehrere Parameter, die gesetzt werden müssen. Dem Parameter *data* wird der verwendete Datensatz übergeben, dem Parameter *y_col* die Spalte für die unabhängige Variable und dem Parameter *d_cols* die Treatment-Variable. Falls, wie im exemplarischen Beispiel, ein Interactive IV Model verwendet wird, muss zusätzlich eine Instrumentalvariable an den Parameter *d_cols* übergeben werden. In der zweiten Zeile wird das Modell ausgewählt, das verwendet werden soll. Im exemplarischen Beispiel wurde das IIVM verwendet. Die zu übergebenden Parameter sind das *DoubleML*-Data Objekt und die zu verwendenden Machine Learner. Zusätzlich kann angegeben werden, ob Crossfitting angewandt werden soll und in wie viele Samples der Datensatz dafür aufgeteilt werden soll. Wann immer eine zufällige Aufteilung der Daten in Samples geschieht, kann ein sogenannter Seed festgelegt werden. Unter Verwendung des gleichen Seeds wird immer dieselbe Aufteilung der Daten erreicht werden. So kann ermöglicht werden, Ergebnisse später zu reproduzieren.

5 Ergebnisse

In diesem Kapitel werden die Ergebnisse der Schätzungen der kausalen Effekte präsentiert und ausgewertet, welche im Zuge der Untersuchung der abhängigen Variablen zur Beantwortung der Leitfragen gewonnen wurden. Jeder Abschnitt dieses Kapitels widmet sich einer der untersuchten abhängigen Variablen. Dabei werden in jedem Abschnitt jeweils zwei Abbildungen dargestellt - eine für jeden Datensatz. Die Ergebnisse werden im Fließtext auf vier und in den Abbildungen auf sechs Nachkommastellen gerundet. Im ersten Abschnitt wird die wahrgenommene Gesundheit der Teilnehmer betrachtet, im zweiten Abschnitt die Anzahl der Arztbesuche innerhalb der letzten sechs Monate und im dritten Abschnitt die Zufriedenheit der Teilnehmer. Die beiden darauf folgenden Abschnitte untersuchen, ob der Bedarf an medizinischen Leistungen und verschreibungspflichtigen Medikamenten gedeckt wurde. Anschließend wird noch die Qualität der erhaltenen medizinischen Leistungen untersucht. Der letzte Abschnitt fasst die Ergebnisse in Bezug auf die wissenschaftliche Literatur zum Thema zusammen und vergleicht sie mit den Ergebnissen anderer Studien zum Oregon Health Insurance Experiment.

5.1 Wahrgenommene Gesundheit

Die Variable **health_gen_12m** gibt den selbstgeschätzten Gesundheitszustand der Teilnehmer wieder - die Skala gab dabei die Optionen „exzellent“, „sehr gut“, „gut“, „okay“ und „schlecht“ her, aus denen die Personen den jeweils zutreffendsten Wert auswählen konnten. Da für die Arbeit mit den Modellen die String-Werte in numerische Werte überführt werden mussten, wurden den einzelnen Ausprägungen Zahlenwerte von 0-4 zugewiesen, wobei 4 für den Gesundheitszustand „exzellent“ und eine 0 für den Zustand „schlecht“ steht. Die Betrachtung dieser Variable sollte dazu beitragen, die Leitfrage, inwieweit MedicAid einen Einfluss auf die Gesundheit der Probanden hat, zu beantworten. Zudem wurde die Variable bereits von anderen Studien untersucht, weshalb die Möglichkeit besteht, Ergebnisse zu vergleichen. Im folgenden werden die Ergebnisse der verschiedenen Durchläufe aufgezeigt:

Bei der Schätzung der kausalen Effekte auf den vollständigen Datensatz (Abbildung 8) ist zu erkennen, dass die Intention to treat einen geringen Koeffizienten hat. Mit einer Höhe von 0,0092 ist der Koeffizient so niedrig, dass der Effekt vernachlässigbar gering zu sein scheint. Wird der Treatment-Effekt geschätzt, so ergeben sich mit einem Interactive Regression Model ein Wert von 0,0282 und mit dem Interactive IV Model ein Wert von 0,0476. Es fällt auf, dass die Wirkungsrichtung beider Effekte positiv ist und beide Effekte auf dem 1%-Niveau statistisch signifikant sind.

Ein Vergleich der vorigen Ergebnisse mit denen aus dem randomisierten Datensatz (Abbildung 9) zeigt, dass die Effekte in allen drei untersuchten Szenarien sehr ähnlich geschätzt wurden. Der Effekt der Intention to treat liegt bei 0,0113, während die Koeffizienten für den Treatment-Effekt vom IRM auf 0,0269 und vom IIVM auf 0,0488 geschätzt wurden. Auch hier sind alle Ergebnisse signifikant, wie am t-Test gesehen werden kann. Da der vollständige Datensatz streng genommen kein randomisiertes Experiment darstellt, sollte die Schätzung des Interactive IV Models am robustesten sein. Bei dem randomisierten Datensatz sollte die Verwendung einer Instrumentalvariable nicht notwendig sein - im Falle von Störfaktoren kann es jedoch trotzdem sinnvoll sein Instrumentalvariablen zu verwenden. An dieser Stelle lässt sich nicht mit Sicherheit sagen, welches Modell den kausalen Effekt besser schätzt weshalb dieser als Intervall zwischen den Schätzungen des IRM und IIVM angegeben wird. Damit liegt der kausale Koeefizient zwischen 0,0269 und 0,0488. Zwar ist die Wirkungsrichtung des Koeffizienten positiv, doch selbst das obere Ende des Intervalls ist mit einem Koeffizienten in Höhe von 0,0488 so gering, dass vermutlich nicht von einem relevanten kausalen Effekt gesprochen werden kann. Es scheint zwar einen Wirkungszusammenhang zwischen dem Treatment und **health_gen_12m** zu geben, doch in Anbetracht der Tatsache, dass die Skala der

health_gen_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.009184	0.003328	2.759999	0.00578	0.002662	0.015707
Treatment Effekt - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.028206	0.003353	8.412796	4.00E-17	0.021635	0.034777
Treatment Effekt - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.047588	0.01606	2.963045	0.003046	0.01611	0.079065

Abbildung 8: Kausale Effekte „Wahrgenommene Gesundheit“ - A (Quelle: E. D.)

health_gen_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.011298	0.003925	2.878413	0.003997	0.003605	0.018992
Treatment Effekt - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.026924	0.004365	6.167512	6.94E-10	0.018368	0.03548
Treatment Effekt - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.048759	0.017578	2.773913	0.005539	0.014307	0.083211

Abbildung 9: Kausale Effekte „Wahrgenommene Gesundheit“ - B (Quelle: E. D.)

Gesundheit von 0-4 reicht, ist eine Verbesserung des Gesundheitszustandes um 0,0488 bei Erhalt des Treatments nicht von Bedeutung.

Im Zuge der Beantwortung der Leitfrage konnte im Rahmen der Untersuchung zwar ein positiver Kausalzusammenhang zwischen MedicAid und wahrgenommener Gesundheit festgestellt werden, jedoch ist dieser vernachlässigbar gering. Es sind viele Szenarien denkbar, wie das Ergebnis der Analyse erklärt werden kann. Beispielsweise bedeutet die Möglichkeit dazu eine Behandlung finanziert zu bekommen nicht automatisch, dass ein Patient diese in Anspruch nimmt. Es wäre denkbar, dass Patienten mit ihren Beschwerden einen Arzt aufsuchen, im Anschluss aber die empfohlenen Therapie-Möglichkeiten nicht wahrnehmen. Der Datensatz des OHIE bietet keine Möglichkeit zu erkennen, ob Patienten tatsächlich Behandlungen bekommen und kontinuierlich bis zum Abschluss wahrgenommen haben. Es sollte also für möglich gehalten werden, dass MedicAid in diesem Fall nicht das Treatment, sondern die Option zum Erhalt eines Treatments ist. Wenn dem so wäre, könnte der gemessene kausale Effekt von MedicAid auf die Gesundheit von Menschen wieder als eine Art „Intention to treat“-Effekt betrachtet werden. Was zusätzlich berücksichtigt werden sollte ist zudem, dass die Beurteilung des Gesundheitszustandes eine Selbsteinschätzung der Probanden war. Es ist also nicht sichergestellt, dass alle Teilnehmer des OHIE das gleiche Verständnis der potenziellen Antwortoptionen zum eigenen Gesundheitszustand hatten, da diese subjektive Kriterien waren. Es ist denkbar, dass diese Abweichungen sich im Mittel ausgleichen, jedoch wäre es für zukünftige Arbeiten sinnvoller ein objektives Kriterium für die Abbildung des Gesundheitszustandes der Probanden zu bilden. Die gerechtfertigtste Interpretation ist an dieser Stelle, dass der Sachverhalt nicht mit Sicherheit korrekt interpretiert werden kann und zusätzliche Daten gesammelt und untersucht werden sollten, um mehr Aufschluss zu geben.

Diese Ergebnisse stimmen zu einem gewissen Grad mit der bisherigen Studienlage zum OHIE überein. Finkelstein et al. schreiben in ihrem Paper „The Oregon Health Insurance Experiment: Evidence from the First Year“, dass versicherte

doc_num_mod_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.011647	0.015533	0.74978	0.453387	-0.018798	0.042091
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.151387	0.015568	9.724253	2.38E-22	0.120875	0.1819
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.078637	0.075736	1.038302	0.299129	-0.069803	0.227078

Abbildung 10: Kausale Effekte „Anzahl Arztbesuche“ - A (Quelle: E. D.)

doc_num_mod_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.021015	0.01946	-1.079902	0.280186	-0.059156	0.017126
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.192995	0.019247	10.027286	1.16E-23	0.155272	0.230719
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.020602	0.086781	0.237405	0.812343	-0.149486	0.190691

Abbildung 11: Kausale Effekte „Anzahl Arztbesuche“ - B (Quelle: E. D.)

Individuen eine um 25% höhere Wahrscheinlichkeit haben, ihren Gesundheitszustand mit „gut“, „sehr gut“ oder „exzellent“ zu bewerten, als die Kontrollgruppe (Finkelstein et al., 2012). Dies deutet auf einen positiven Zusammenhang zwischen Versicherung und guter Gesundheit hin. Baicker et al. hingegen fanden in ihrer Publikation „The Oregon Experiment — Effects of MedicAid on Clinical Outcomes“ keinen signifikanten Effekt von MedicAid auf die physische Gesundheit (Baicker et al., 2013). Die Ergebnisse dieser Arbeit stimmen hinsichtlich der Wirkungsrichtung mit dem von Finkelstein et al. ermittelten Effekt überein, schätzen diesen jedoch deutlich geringer ein. Der durch Double Machine Learning ermittelte kausale Effekt in Höhe von 0,0269-0,0488 wurde als vernachlässigbar gering eingeschätzt, was mehr mit den Ergebnissen von Baicker et al. übereinstimmt. Es kann angenommen werden, dass die Ergebnisse unter Anwendung von Double Machine Learning eine höhere Validität aufweisen, da diese den tatsächlichen kausalen Effekt ermitteln, anstatt Assoziationen zu belegen.

5.2 Anzahl der Arztbesuche

Die Variable **doc_num_mod_12m** steht für die Anzahl an Arztbesuchen, welche die Teilnehmer innerhalb der letzten sechs Monate getätigt haben. Diese Variable lag bereits in numerischer Form vor, also war kein Data-Cleaning neben dem Umgang mit Missing Values nötig. Im Folgenden sind die Ergebnisse der Untersuchung:

Bei der Schätzung des kausalen Effekts von MedicAid auf die Anzahl an Arztbesuchen ist erkennbar, dass die Ergebnisse der Intention to treat und des IIVM nicht statistisch signifikant sind, da der t-Test das 1%-Signifikanzniveau überschreitet (Abbildung 10). Lediglich der Koeffizient des IRM in Höhe von 0,1514 ist signifikant. Die Ergebnisse des vollständigen Datensatzes weisen also auf einen positiven Zusammenhang zwischen Treatment und Outcome hin. Unter Betrachtung der Ergebnisse auf dem randomisierten Datensatzes (Abbildung 11) fällt auf, dass wieder lediglich die Schätzung des Interactive Regression Models mit einem Koeffizienten von 0,193 signifikant ist. Da die anderen Koeffizienten das 1%-Signifikanzniveau nicht unterschreiten, werden diese nicht betrachtet. Sowohl bei dem vollständigen Daten-

satz, als auch bei dem auf **numhh** bedingten Datensatz, ist die Wirkungsrichtung der signifikanten Koeffizienten positiv.

An den Ergebnissen ist erkennbar, dass ein positiver kausaler Effekt vorliegt von dem Treatment auf die Anzahl der Arztbesuche vorliegt. Zwar sind die Ergebnisse zu der Intention to treat und die des IIVM nicht signifikant, jedoch haben die Schätzungen des Interactive Regression Models auf beiden Datensätzen signifikante Schätzungen hervorgebracht. Da das IRM nicht dafür geeignet ist, mit Nicht-randomisierten Daten zu arbeiten, wird die Schätzung des Koeffizienten auf dem randomisierten Datensatz als zuverlässiger betrachtet (Bach et al., 2021). Es wird an dieser Stelle also interpretiert, dass der kausale Koeffizient von MedicAid auf die Anzahl an Arztbesuchen bei 0,193 liegt.

Demnach lässt sich interpretieren, dass MedicAid-Coverage dazu führte, dass die Teilnehmer der Studie öfter Ärzte aufgesucht haben. Eine Abdeckung durch MedicAid erhöht durchschnittlich die Anzahl an Arztbesuchen um etwa 0,193 mal. Während dieser Effekt für ein Individuum isoliert betrachtet gering zu sein scheint, ist zu berücksichtigen, dass der Effekt über viele Personen hinweg ins Gewicht fallen kann. Geht jede von 100 Personen im Schnitt zusätzliche 0,193 mal zum Arzt, sind das etwa 19 zusätzliche Arztbesuche. Eine mögliche Erklärung dafür ist, dass ein Arztbesuch ohne Versicherung in der Regel mit Kosten verbunden ist. Gerade in den USA fallen Kosten für medizinische Leistungen sehr hoch aus, weshalb potenziell der Trade-off zwischen Kosten und Dringlichkeit einer Behandlung beziehungsweise eines Arztbesuches abgewogen wird (Vladeck, 2003). Durch die Abdeckung eines MedicAid-Programms, welches im Falle eines Besuches oder einer medizinischen Prozedur die Kosten trägt, muss dieser Trade-off nicht mehr gemacht werden - die wirtschaftliche Hemmschwelle einen Arzt zu besuchen könnte also abnehmen.

Diese Ergebnisse decken sich mit denen von Baicker et al., welche zeigen, dass versicherte Individuen häufiger Ärzte besuchen, als die Kontrollgruppe. Sie fanden einen Unterschied im Mittelwert zwischen Versuchs- und Kontrollgruppe von 2,7 (Baicker et al., 2013). Der Effekt konnte auch in dieser Arbeit belegt werden, wenn auch deutlich geringer (0,193). Auch an dieser Stelle ist davon auszugehen, dass die Ergebnisse unter Verwendung von Double Machine Learning robuster und somit zuverlässiger sind.

5.3 Zufriedenheit

In der Variable **happiness_12m** wurde die selbst wahrgenommene Zufriedenheit der Teilnehmer erfasst. Ähnlich wie bei der Variable zum Gesundheitszustand gab es hier ebenfalls eine Item-Skala, bei der die Teilnehmer ihre Zufriedenheit mit den Attributen „sehr glücklich“, „glücklich“ und „nicht glücklich“ bewerten konnten. Auch hier wurden die Strings im Preprocessing in numerische Werte, in diesem Fall von 0 bis 2 überführt, wobei 2 für „sehr glücklich“ steht. Die Grundzufriedenheit der Bevölkerung ist eine wichtige Variable, die sich auf viele weitere Faktoren auswirkt. Generell gilt, dass eine höhere Zufriedenheit der Menschen sowohl auf politischer und gesellschaftlicher, als auch auf individueller Ebene angestrebt wird (Malthus, 1888). Im Folgenden sind die Ergebnisse der Untersuchung aufgeführt:

Die Koeffizienten zum kausalen Effekt von MedicAid auf die Zufriedenheit der Teilnehmer sind im ersten Datensatz allesamt signifikant (Abbildung 12). Der Effekt der Intention to treat und des IRM sind mit 0,0212 und 0,0322 ähnlich hoch. Der vom Interactive IV Model geschätzte Koeffizient liegt bei 0,1227 und ist damit deutlich höher.

Die Ergebnisse ähneln den Ergebnissen vom randomisierten Datensatz (Abbildung 13). Auch hier sind alle Koeffizienten signifikant. Die Intention to treat hat einen Koeffizienten von 0,0292 und IRM und IIVM haben einen Koeffizienten von 0,0292 bzw. 0,1189. Alle kausalen Effekte haben dabei eine positive Wirkungsrichtung.

Auf dem vollständigen Datensatz haben die Schätzungen des IIVM die höchste theoretische Validität. Damit lässt sich die Obergrenze des kausalen Effektes auf 0,1227 schätzen. Die Schätzung des IRM ist nur auf dem randomisierten Datensatz valide, wo sie sich jedoch nicht signifikant von der Schätzung der Intention to treat unterscheidet. Wie zuvor erwähnt, kann nur von einem kausalen Effekt gesprochen werden, wenn der geschätzte Effekt und die Intention to treat verschieden voneinander sind. Auch hier wird wieder ein Intervall für den kausalen Koeffizienten geschätzt. Dieser liegt zwischen 0,0292 und 0,1227, wobei die untere Intervallgrenze nahezu identisch mit der Intention to treat ist. Anhand des Intervalls lässt sich kein relevanter Effekt bzw. ein positiver Effekt von MedicAid-Coverage auf die Zufriedenheit der Empfänger interpretieren. Im Verhältnis zur Größe der Skala ist dieser jedoch gering - der Koeffizient macht selbst bei Verwendung der oberen Intervallgrenze nur etwa ein Achtel des Sprunges zur nächsten Stufe aus. Theoretisch wäre denkbar gewesen, dass die Sorge um medizinische Not-Situationen und potenziell hohe Kosten die Zufriedenheit der Nicht-Versicherten mindern beziehungsweise die Zufriedenheit der Versicherten durch die Auflösung dieser Sorge steigern - diese Vermutung lässt sich empirisch anhand der Daten des Oregon Health Insurance Experiments zumindest teilweise bestätigen beziehungsweise nicht vollständig widerlegen. Ob weniger Sorge um medizinische Kosten tatsächlich der Grund ist, oder noch ein anderer Aspekt dahintersteht, müsste in Zukunft näher untersucht werden.

Der Einfluss von MedicAid auf die Zufriedenheit wurde auch von Baicker et al. untersucht. Diese fanden keinen signifikanten Effekt von MedicAid auf die Zufriedenheit der Versuchsgruppe (Baicker et al., 2013). Da im Rahmen dieser Arbeit ein Intervall für den kausalen Effekt geschätzt wurde, ist der Vergleich nur begrenzt möglich. Unter Verwendung der unteren Intervallgrenze von 0,0292 lässt sich der Effekt als vernachlässigbar gering einschätzen und stimmt somit mit den Ergebnissen von Baicker et al. überein. Die Ver-

happiness_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.021169	0.003375	6.273152	3.54E-10	0.014555	0.027783
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.032173	0.003365	9.560503	1.17E-21	0.025578	0.038769
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.122665	0.016633	7.374804	1.65E-13	0.090065	0.155265

Abbildung 12: Kausale Effekte „Zufriedenheit“ - A (Quelle: E. D.)

happiness_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.029178	0.003969	7.350671	1.97E-13	0.021398	0.036958
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.029196	0.003971	7.353267	1.93E-13	0.021414	0.036978
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.118856	0.01778	6.684848	2.31E-11	0.084008	0.153704

Abbildung 13: Kausale Effekte „Zufriedenheit“ - B (Quelle: E. D.)

wendung der oberen Intervallgrenze von 0,1227 hingegen deutet auf einen kausalen Zusammenhang zwischen Treatment und Outcome hin, wenn auch einen sehr geringen. An dieser Stelle wird darauf verwiesen, dass zukünftig weitere Untersuchungen bzw. Studien den Zusammenhang zwischen Gesundheitsversicherung und Zufriedenheit untersuchen sollten, bevor eindeutige Schlussfolgerungen gezogen werden können.

5.4 Medizinischer Bedarf gedeckt

Ob ein Teilnehmer sämtliche, innerhalb der letzten sechs Monate benötigten, medizinischen Leistungen erhalten hat, wurde in der Variable **needmet_med_cor_12m** erfasst. Die Befragten konnten angeben, dass sie entweder keine Leistungen benötigten, oder diese alle erhalten haben - in diesem Fall steht in der bereinigten Spalte eine 1. Hat der Teilnehmer nicht alle benötigten medizinischen Leistungen erhalten, ist dies durch eine 0 gekennzeichnet. Die Frage danach, ob sämtlicher Bedarf an medizinischen Leistungen einer Person gedeckt werden kann, ist eine wichtige. Sind bestimmte medizinische Leistungen für manche Menschen aufgrund ihrer finanziellen Situation nicht erschwinglich, dann stellt dies ein ethisches Problem dar (Summers & Morrison, 2009). Ob die Kosten medizinischer Leistungen tatsächlich einen Einfluss darauf haben, ob diese bei Bedarf in Anspruch genommen werden, sollte untersucht werden. Nachfolgend sind die Ergebnisse der Analyse abgebildet:

Die geschätzten kausalen Koeffizienten, die MedicAid auf den Erhalt benötigter medizinischer Leistungen hat, sind im vollständigen Datensatz allesamt nicht signifikant (Abbildung 14). Die Ergebnisse im bedingten Datensatz sind ebenfalls nicht signifikant und werden daher nicht weiter betrachtet (Abbildung 15).

An dieser Stelle lässt sich lediglich die Annahme äußern, dass Behandlungen, welche tatsächlich notwendig sind, von den Betroffenen Personen trotz hohen Kosten dennoch in Anspruch genommen werden. Theoretisch lässt sich hier auch ein Fall von Survivorship Bias vermuten - Personen die wirklich notwendige Behandlungen trotz hoher Kosten in Anspruch nehmen, haben vermutlich eine höhere Chance im Datensatz aufzutauchen, als Menschen, die dringende Behandlungen aus finanziellen Gründen nicht in Anspruch nehmen (können), da verstorbene Personen aus dem Datensatz entfernt wurden. Dies sind jedoch nur Vermutungen, welche

needmet_med_cor_12m_Yes						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000045	0.000044	1.014984	0.310114	-0.000042	0.000131
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.00015	0.000094	-1.596246	0.110434	-0.000334	0.000034
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000207	0.000295	0.701115	0.483231	-0.000371	0.000785

Abbildung 14: Kausale Effekte „Bedarf Med. Leistungen“ - A (Quelle: E. D.)

needmet_med_cor_12m_Yes						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000072	0.000055	1.305576	0.191697	-0.000036	0.000181
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000212	0.000125	1.702199	0.088718	-0.000032	0.000457
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000188	0.000217	0.867655	0.385583	-0.000237	0.000614

Abbildung 15: Kausale Effekte „Bedarf Med. Leistungen“ - B (Quelle: E. D.)

an dieser Stelle nicht mit empirischen Belegen aus den Ergebnissen dieser Arbeit unterstützt werden können.

Ob MedicAid einen Einfluss darauf hat, inwieweit der Bedarf an medizinischen Leistungen gedeckt wird, wurde bereits von Baicker et al. im Paper „The Effect of MedicAid on Management of Depression“ untersucht. Sie fanden, dass durch MedicAid die Wahrscheinlichkeit, dass der Bedarf an medizinischen Leistungen nicht gedeckt wird, um 10,7% reduziert wurde (Baicker et al., 2018). Diese Ergebnisse können anhand der Untersuchung dieser Arbeit weder empirisch belegt, noch widerlegt werden. Es wäre in Zukunft interessant, ähnliche Daten und damit auch diese Fragestellung, genauer zu untersuchen.

5.5 Bedarf an Medikamenten gedeckt

Die Variable **needmet_rx_cor_12m** misst, ob der Bedarf eines Studienteilnehmers an verschreibungspflichtigen Medikamenten innerhalb der letzten sechs Monate gedeckt wurde. Auch hier steht eine 1 dafür, dass keine Medikamente benötigt oder alle benötigten Medikamente erhalten wurden, während eine 0 angibt, dass nicht alle Bedürfnisse gedeckt wurden. Das Konzept dieser Fragestellung ähnelt der des vorherigen Unterabschnitts, jedoch werden medizinische Behandlungen und Medikamente im Rahmen dieser Arbeit als zwei unterschiedliche Ausprägungen betrachtet. Beispielsweise können die Kosten für die Behandlung chronischer Erkrankungen, die eine langfristige medikamentöse Therapie erfordern, im Vergleich zu den einmaligen Kosten einer Operation erheblich höher und fortlaufend sein (Hussey et al., 2014). Diese potenzielle Unterscheidung war Grund dafür die Variable **needmet_rx_cor_12m** separat zu untersuchen. Die Ergebnisse sind im Folgenden zu sehen:

Bei der Untersuchung des vollständigen Datensatzes ist lediglich der Koeffizient des Interactive Regression Models signifikant (Abbildung 16). Dieser beträgt -0,0003 und ist zwar negativ, aber kann approximativ als 0 betrachtet und vernachlässigt werden. Bei der Schätzung auf dem vollständigen Datensatz ist das IRM weniger geeignet, als das IIVM, allerdings kann die generelle Größenordnung des Koeffizienten angenommen werden.

Bei den Ergebnissen auf dem randomisierten Datensatz ist erneut nur die Schätzung des Interactive Regression Mo-

needmet_rx_cor_12m_Yes						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000084	0.000052	1.629699	0.103165	-0.000017	0.000185
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.000321	0.000114	-2.810998	0.004939	-0.000544	-0.000097
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.00028	0.000307	0.913443	0.36101	-0.000321	0.000881

Abbildung 16: Kausale Effekte „Bedarf Medikamente“ - A (Quelle: E. D.)

needmet_rx_cor_12m_Yes						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.000049	0.000074	-0.656675	0.51139	-0.000193	0.000096
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.001031	0.000215	-4.788631	0.000002	-0.001453	-0.000609
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.000301	0.000334	-0.901043	0.367566	-0.000955	0.000354

Abbildung 17: Kausale Effekte „Bedarf Medikamente“ - B (Quelle: E. D.)

dels signifikant (Abbildung 17). Der geschätzte Koeffizient beträgt hier -0,001 und ist ebenfalls negativ und approximativ 0. An dieser Stelle lässt sich interpretieren, dass MedicAid-Coverage im Rahmen der Daten des Oregon Health Insurance Experiments keinen messbaren kausalen Effekt auf die Abdeckung des Bedarfs an Medikamenten hatte. Es lässt sich vermuten, dass notwendige Medikamente unabhängig von den finanziellen Aspekten in Anspruch genommen werden. An dieser Stelle wird jedoch ausdrücklich darauf hingewiesen, dass die empirische Evidenz in diesem Abschnitt begrenzt ist und weitere Untersuchungen notwendig sind, um diese Hypothese zu bestätigen. Im Zuge dessen wäre es interessant, weitere Variablen wie z.B. finanzielle Aspekte in die Betrachtung einzuziehen.

Bisher wurde noch keine wissenschaftliche Untersuchung darüber durchgeführt, inwiefern MedicAid den Bedarf der Empfänger an verschreibungspflichtigen Medikamenten deckt. Deshalb ist ein direkter Vergleich mit bisheriger empirischer Evidenz nicht möglich. Baicker et al. publizierten jedoch in ihrem Paper „The Effect of MedicAid on Management of Depression“, dass MedicAid zu einem erhöhten Einsatz von Medikamenten bei der Kontrollgruppe führte (Baicker et al., 2018). Theoretisch wäre denkbar, dass ein vermehrter Einsatz von Medikamenten auch bedeutet, dass der Bedarf an Medikamenten besser gedeckt wird. Andererseits ist ebenso vorstellbar, dass der Mehr-Einsatz von Medikamenten kaum Überschneidung mit dem tatsächlichen Bedarf hat und überwiegend „optionale“ Medikamente verwendet werden. Da die Ergebnisse dieser Arbeit die Kausalbeziehung nicht näher beleuchten, bedarf es in Zukunft weiterer Untersuchungen, um den Zusammenhang zu klären.

5.6 Qualität medizinischer Leistungen

Die letzte untersuchte Variable ist **med_qual_12m**. Diese Variable erfasst, inwieweit die Teilnehmer die Qualität der in den letzten sechs Monaten in Anspruch genommenen medizinischen Leistungen beurteilen. Hier wurde ebenfalls wieder eine Item-Skala verwendet, welche im Preprocessing in numerische Form umgewandelt wurde. Die Skala bestand aus den Attributen „exzellent“, „sehr gut“, „gut“, „in Ordnung“, „schlecht“ und „keine Leistungen beansprucht“. Der Skala wurden die Werte 0 bis 4 vergeben,

med_qual_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.007299	0.005358	-1.362208	0.173132	-0.0178	0.003203
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.003905	0.005393	-0.724185	0.468952	-0.014475	0.006664
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.024027	0.02328	-1.032073	0.302038	-0.069655	0.021602

Abbildung 18: Kausale Effekte „Qualität medizinischer Leistungen“ - A (Quelle: E. D.)

med_qual_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.014055	0.006442	-2.18187	0.029119	-0.026681	-0.001429
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.001862	0.00634	0.293614	0.769053	-0.010565	0.014289
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.052689	0.025323	-2.080666	0.037465	-0.102322	-0.003057

Abbildung 19: Kausale Effekte „Qualität medizinischer Leistungen“ - B (Quelle: E. D.)

wobei „exzellent“ als Wert 4 gesetzt worden ist. Wenn die Variable **med_qual_12m** als Kovariable verwendet wurde, wurde „keine Leistungen beansprucht“ mit dem Mittelwert aufgefüllt - damit der potenziell vorhandene Erklärungsgehalt jedoch nicht verloren geht, wurde in einer neu erzeugten binären Spalte festgehalten, welche Teilnehmer keine Leistungen in Anspruch genommen haben. Für die Untersuchung innerhalb dieses Abschnitts wurden zusätzlich zu allen fehlenden Werten in **med_qual_12m** zudem auch die Teilnehmer aus dem Datensatz entfernt, die keine medizinischen Leistungen innerhalb der letzten sechs Monate beansprucht haben. Damit soll sichergestellt werden, dass kein Bias eingeführt wird.

Die Ergebnisse auf dem vollständigen Datensatz sind allesamt nicht auf dem 1%-Niveau signifikant und werden daher nicht weiter betrachtet (Abbildung 18). Die Schätzungen auf dem randomisierten Datensatz sind, ebenfalls nicht auf dem 1%-Niveau signifikant (Abbildung 19). Die Intention to treat und das Interactive IV Model sind jedoch auf dem 5%-Niveau statistisch signifikant. Der Koeffizient der Intention to treat beträgt -0,0141 und der des IIVM beträgt -0,0527. Damit ist der geschätzte Effekt von der Intention to treat verschieden - es scheint also einen negativen kausalen Zusammenhang zwischen MedicAid und der Qualität der erhaltenen medizinischen Leistungen zu geben. In Anbetracht der Größe der Skala lässt sich jedoch wieder feststellen, dass der Effekt vernachlässigbar gering ist. Ebenfalls ist hier anzumerken, dass wie bereits in Kapitel 5.1 mit einer subjektiven Skala gearbeitet wurde. Insbesondere auf die Bewertung von medizinischen Leistungen können viele subjektive Faktoren einwirken, die nicht zwangsläufig mit der Qualität der Leistung zusammenhängen. Ebenfalls sollte angemerkt werden, dass der durchschnittlichen Bevölkerung die fachlichen Qualifikationen fehlen könnten, um aussagekräftige Bewertungen medizinischer Leistungen treffen zu können.

Im Rahmen dieser Arbeit lässt sich also interpretieren, dass MedicAid sich nicht merklich auf die Qualität der erhaltenen medizinischen Leistungen auswirkt, die empirische Evidenz jedoch gering ist. Aus ethischen Gründen sollte jeder Patient unabhängig von seiner Versicherung bestmöglich behandelt werden (Summers & Morrison, 2009). Es wäre jedoch von großem Interesse zu untersuchen, ob diese Idealvorstellung in der Realität umgesetzt wird. Die Varia-

ble **med_qual_12m** wurde bislang in keiner anderen Publikation zum Oregon Health Insurance Experiment untersucht, was keinen Vergleich empirischer Evidenz ermöglicht.

5.7 Einordnung OHIE

Die Menge der statistisch signifikanten Ergebnisse, welche aus dieser Arbeit hervorgehen, ist limitiert. Eine direkte Überschneidung der Ergebnisse mit denen anderer Studien zum OHIE besteht nur bei drei untersuchten Variablen (siehe Kapitel 5.1-5.3). Ein Vergleich dieser Ergebnisse mit anderen Studien zeigt, dass es sowohl Übereinstimmungen als auch Unterschiede gibt. Die Übereinstimmung gibt es primär in der Wirkungsrichtung der geschätzten Effekte. Unterschiede gibt es jedoch in ihrer Größe - es zeigte sich (in der limitierten Samplegröße $n = 3$) dass Double Machine Learning tendenziell geringere Effektgrößen geschätzt hat, als klassische statistische Methoden. Eine Einordnung in die bisherigen Ergebnisse zum Oregon Health Insurance Experiment lässt sich wie folgt formulieren: Die Untersuchungen kausaler Effekte durch MedicAid auf verschiedene Variablen haben gezeigt, dass andere Studien zum OHIE Kausalzusammenhänge in ihrer Wirkungsrichtung akkurat beleuchten, jedoch in der Größe der Effekte nicht akkurat bestimmen konnten. Da im Rahmen des OHIE viele weitere Faktoren und Variablen untersucht wurden, wäre es interessant diese Studien ebenfalls mit Double Machine Learning zu untersuchen und potenziell zu reevaluieren. Gerade wenn die Ergebnisse des Oregon Health Insurance Experiments Basis für zukünftige Entscheidungen zum Thema Gesundheitsversicherung bilden könnten, sollte aus wissenschaftlichem und generellem Interesse sichergestellt werden, dass die zugrundeliegenden Ergebnisse so akkurat wie möglich sind. Zukünftige Untersuchungen mit Double Machine Learning auf dem Datensatz des Oregon Health Insurance Experiments könnten beispielsweise zur mentalen Gesundheit der Teilnehmer durchgeführt werden - in diesem Bereich liegen bereits ausführliche Ergebnisse durch Baicker et al. vor, welche signifikante Effekte von MedicAid auf die Diagnose und Behandlung von Depressionen fanden (Baicker et al., 2018).

6 Kritik

Gerade bei der Auswertung von Daten, sollte jeder Schritt kritisch hinterfragt werden. Kleinste Fehler in der Datenerhebung, bei der Modellwahl oder bei anderen Prozessen, können einen Bias in das Ergebnis einführen und dieses damit verzerren. Laufen Prozesse beispielsweise, den äußeren Umständen geschuldet, nicht optimal ab, so ist es wichtig, darauf hinzuweisen. Dadurch soll verhindert werden, dass zukünftige Arbeiten Bezug auf die Ergebnisse nehmen und unbewusst den Bias weiterreichen. Zudem kann durch transparentes Arbeiten ein Bewusstsein für die bestimmte Arten von Fehlerquellen geschärft werden. In diesem Kapitel soll der gesamte Prozess, von Erhebung der Daten, bis hin zur Gewinnung der Ergebnisse kritisch betrachtet und auf potenzielle Fehlerquellen untersucht werden. Im folgenden Unterkapitel wird dabei zunächst der Datensatz des Oregon Health Insurance Experiments auf mögliche Fehlerquellen untersucht. Im zweiten Abschnitt wird die statistische Signifikanz der Ergebnisse betrachtet. Im dritten und letzten Abschnitt wird auf die Limitationen dieser Arbeit hingewiesen.

6.1 Datensatz

Während das OHIE ein großes Experiment von hoher Relevanz ist, lassen sich jedoch einige Kritikpunkte am Datensatz äußern. Im Rahmen dieser Arbeit werden die folgenden fünf Kritikpunkte am OHIE als die wichtigsten erachtet:

- Die Datensätze enthalten nicht genug Beobachtungen
- Es liegt potenziell ein Sampling Bias vor
- Es werden nicht alle Variablen von Interesse betrachtet
- Variablen unterliegen teilweise Subjektivität
- Die Daten spiegeln nur eine Momentaufnahme wieder

Ein wesentlicher Kritikpunkt ist die Samplegröße der jeweiligen Umfragen. Die Umfrage nach 12 Monaten, welche die Grundlage für die Analysen in dieser Arbeit war, enthielt nach dem Preprocessing noch 22.904 Datenpunkte im vollständigen und 16.064 Datenpunkte im randomisierten Datensatz. Organisatorische und finanzielle Gründe waren vermutlich limitierende Faktoren bei der Entscheidung der Samplegrößen der Umfragen, jedoch wäre es aus Datenperspektive besser gewesen, die Grundgesamtheit der Teilnehmer zu befragen.

Zusätzlich muss davon ausgegangen werden, dass es potenziell mehrere Quellen für einen Sampling Bias gibt. Im Falle des Oregon Health Insurance Experiments war die Auswahl der untersuchten Gesamtheit nicht zufällig - es wurden ausschließlich Menschen mit Wohnsitz in Oregon in das Sample aufgenommen. Während diese Tatsache in der Realität vermutlich keinen signifikanten Bias auf die Ergebnisse überträgt, sollte dennoch beachtet werden, dass theoretisch keine vollständige Übertragbarkeit der Ergebnisse gewährleistet wäre, falls sich Oregon signifikant von anderen Staaten in den USA unterscheidet. Wäre beispielsweise die medizinische Versorgung in Oregon signifikant schlechter, als in dem Rest Amerikas, würde das voraussichtlich kausale Effekte der öffentlichen Krankenversicherung auf die Gesundheit der Bevölkerung in ihrer Stärke beeinflussen oder möglicherweise sogar weitere Variablen durch Confounding betreffen. Gleiches gilt für die Übertragbarkeit auf andere Länder. Ebenfalls zu beachten ist, dass eine große Menge an Teilnehmern die Umfragen nicht beantwortet haben. Unterscheidet sich die Gruppe der Nicht-Antwortenden von den Teilnehmern, welche die Umfragen beantwortet haben, liegt ebenfalls ein Sampling Bias vor. Es wurden bereits monetäre Anreize gesetzt und Nachbearbeitungen durchgeführt, um die Antwortquote zu erhöhen, jedoch hätten diese potenziell höher bzw. intensiver sein müssen. In dieser Arbeit wird jedoch angenommen, dass der Effekt vernachlässigbar gering ist.

Ebenfalls kritisieren lässt sich, dass Variablen von potenziellem Interesse nicht im Datensatz erfasst wurden. Wie zuvor erwähnt besteht beispielsweise keine Möglichkeit in den Daten zu erkennen, ob Patienten tatsächlich Behandlungen bekommen und kontinuierlich bis zum Abschluss wahrgenommen haben. Gerade im Bereich der Kausalen Inferenz ist es von hoher Wichtigkeit, dass für den Kausalzusammenhang relevante Variablen nicht ausgelassen werden, da dies einen Bias einführen kann (Hernan & Robins, 2023). Zukünftig wäre es interessant mehr Variablen zu finanziellen Aspekten in den Datensatz einzubeziehen, da die Übernahme von Kosten ein wesentlicher Unterschied zwischen öffentlicher und privater Krankenversicherung ist (White, 2015).

Neben ausgelassenen Variablen lassen sich auch die erfassten Variablen kritisieren. Viele der Variablen wurden über subjektive Selbsteinschätzungen erhoben. Neben möglichen Faktoren, welche die Selbsteinschätzungen verzerrt haben könnten, ist es zudem denkbar, dass die Ausprägungen der Variable von den Teilnehmern unterschiedlich interpretiert wurden: was ein Proband als „exzellenter Gesundheitszustand“ einschätzt, ist immer relativ zur eigenen Wahrnehmung, nicht zu einem objektiven Standard. Für zukünftige Studien wäre es hier potenziell sinnvoll, zusätzlich zu subjektiven Variablen auch objektive Kenngrößen zu messen - im Falle der Gesundheit ließen sich beispielsweise Puls, Blutdruck, Sauerstoffsättigung, Gewicht, etc. ohne großen Mehraufwand ebenfalls bestimmen.

Zuletzt kann kritisiert werden, dass die Umfragen immer nur eine Momentaufnahme abbilden können. Zwar kann durch die Kombination der Umfragen ein längerer Zeitraum abgedeckt werden, jedoch lässt sich argumentieren, dass gerade bei langwierigen Themen wie Medizin und Gesundheit ein weitreichenderer Zeitraum hätte betrachtet werden müssen. Zusätzlich ist zu berücksichtigen, dass aufgrund der geringen Anzahl an Umfragen ein höheres Risiko besteht, dass die Umfrageergebnisse (insbesondere die subjektiven) durch äußere Umstände beeinflusst werden. Dadurch könnte die interne Validität der Studie eingeschränkt sein. Eine Erhöhung der Anzahl der Beobachtungen würde die Robustheit der Ergebnisse erhöhen und das Risiko von Verzerrungen durch äußere Einflüsse reduzieren (Hamaker et al., 2015).

Es lässt sich an dieser Stelle zusammenfassend sagen, dass das Oregon Health Insurance Experiment einen Datensatz hervorgebracht hat, aus dem sich erste Inferenzen bilden lassen - für eine vollständige Beleuchtung der Thematik sollte jedoch ein ähnliches Experiment erhoben werden, bei denen die potenziellen Fehler des OHIE ausgebessert werden.

6.2 Signifikanz der Ergebnisse

Die in dieser Arbeit gewonnenen Ergebnisse sind an vielen Stellen nicht statistisch auf dem 1%-Niveau (bzw. 5%-Niveau) signifikant gewesen. Während Double Machine Learning eine wertvolle Methode ist, um Kausale Inferenz anzuwenden, ist es nicht garantiert, dass es in allen Fällen statistisch signifikante Ergebnisse liefert. Es gibt zahlreiche Faktoren, die die statistische Aussagekraft der Ergebnisse beeinflusst haben könnten - unter anderem der Stichprobenumfang, das Preprocessing der Daten oder die Komplexität des zugrundeliegenden Kausalzusammenhangs (Croxton & Cowden, 1939; Hernan & Robins, 2023). Eine Wiederholung der Analyse mit anderer Aufteilung des Datensatzes (d.h. Random-Seed bei der Crossvalidation) führt cet. par. potenziell einen Survivorship Bias in die Ergebnisse ein, da lediglich gewünschte Ergebnisse betrachtet werden, weshalb davon in dieser Arbeit abgesehen wurde (Ioannidis, 2005). Generell ist davon abzuraten, statistisch nicht-signifikante Analysen zu wiederholen. Stattdessen sollten die angewandte Methodik und der zugrundeliegende Sachverhalt tiefergehend auf mögliche Fehlerquellen untersucht werden. Eine nicht-exklusive Liste der potenziellen Änderungen an den im Rahmen dieser Arbeit durchgeführten Untersuchungen beinhalten:

- anders mit Missing Values verfahren
- andere Machine Learning Algorithmen verwenden
- zusätzliches Hyperparamter-Tuning
- andere Variablen untersuchen
- Folds (n_folds) bei der Crossvalidation anders stratifizieren

Viele dieser Änderungen sind mit dem in dieser Arbeit bereitgestellten Code mit geringem Aufwand durchführbar. Für die Zukunft wäre es interessant zu sehen, ob die Signifikanz der Ergebnisse von der verwendeten Methodik beeinflusst wurde - im besten Fall lassen sich so langfristig „Best Practices“ im Umgang mit Double Machine Learning ermitteln, welche zukünftige Anwendungen in ihrer Genauigkeit verbessern könnten.

6.3 Limitationen

Eine Limitation bei der Erstellung dieser Arbeit war die benötigte Rechenleistung. Die Anwendung von Machine Learning und der Umgang mit großen Datenmengen ist meist rechnerisch anspruchsvoll und erfordert leistungsfähige Hardware. Für diese Arbeit wurde ein Cloud-Server verwendet, um die hohen Anforderungen für das Errechnen von insgesamt 36 kausalen Effekte (jeweils sechs Modelle für insgesamt sechs Variablen) zu decken. Ein vollständiger Durchlauf des Codes auf der verwendeten Hardware dauerte etwa 8 Stunden - trotz geringem Hyperparameter-Tuning. Ohne diese Limitationen wäre es interessant gewesen, zusätzliche Variablen aus dem Oregon Health Insurance Experiment betrachten zu können. Im Rahmen von weiteren Studien zum OHIE wurde eine Vielzahl von Variablen untersucht, die in dieser Arbeit aufgrund der technischen Limitationen im vorgegebenen Zeitraum nicht betrachtet werden konnten.

7 Schlussbetrachtung

Nachdem im Rahmen dieser Arbeit der Datensatz des Oregon Health Insurance Experiments mithilfe von Double Machine Learning auf kausale Effekte untersucht wurde, lassen sich verschiedene Punkte zusammenfassend betrachten. Zum einen lässt sich sagen, dass das OHIE einen wichtigen Beitrag zur Evaluierung der Auswirkungen von MedicAid-Programmen und öffentlichen Gesundheitssystemen auf verschiedene Faktoren liefern kann. Zum anderen lässt sich die Leitfrage, ob kausale Effekte durch das Treatment MedicAid vorliegen beantworten - ja, die Aufnahme in ein MedicAid-Programm hat kausale Effekte auf verschiedene Variablen. Inwieweit abhängige Variablen durch das Treatment beeinflusst werden, wurde in den vorigen Kapiteln ausgewertet. Damit lassen sich die in der Einführung formulierten Leitfragen teilweise beantworten:

- MedicAid hatte einen vernachlässigbar geringen positiven Einfluss auf die selbsteingeschätzte Gesundheit der Begünstigten (0,0269-0,0488)
- MedicAid hatte einen positiven kausalen Effekt auf die Anzahl der Arztbesuche (0,193)
- MedicAid hatte keinen oder einen nur sehr geringen (positiven) Effekt auf die Zufriedenheit der Teilnehmer gehabt (0,0292-0,1227)
- MedicAid hatte keinen Effekt darauf, ob alle benötigten Medikamente erhalten wurden
- MedicAid hatte einen vernachlässigbar geringen negativen Einfluss auf die Qualität der erhaltenen medizinischen Leistungen

Nicht beantworten lassen sich mit den Ergebnissen dieser Arbeit, ob Menschen ohne MedicAid-Coverage auf eigentlich notwendige medizinische Leistungen verzichten. Dennoch hat sich die Anwendung von Double Machine Learning als größtenteils erfolgreich erwiesen - es wurde jedoch auch ersichtlich, dass ein geeigneter Datensatz von hinreichender Größe notwendig ist, um signifikante Ergebnisse zu erhalten. Zudem hat sich gezeigt, dass es in der Anwendung der verschiedenen Modelle durchaus zu Unterschieden gekommen ist. Aus diesem Grund erscheint es sinnvoll, auch in zukünftigen Anwendungen von Double Machine Learning weiterhin verschiedene Modelle gegeneinander zu testen und den kausalen Effekt konservativ als ein Intervall zu schätzen. Solange nicht die Annahme besteht, dass im Datensatz des OHIE wesentliche Confounder ausgelassen wurden oder der Umgang mit Missing Values den Datensatz signifikant verzerrt hat, ist davon auszugehen, dass die Ergebnisse dieser Arbeit valide und robust sind.

8 Ausblick

Ein wesentliches Ziel dieser Arbeit, neben der Beantwortung der Leitfragen, war eine Grundlage für zukünftige Literatur und Forschung zu schaffen. Kausale Inferenz ist eine noch junge akademische Disziplin, die eine hohe Einstiegsbarriere hat. Da Statistik jedoch unerlässlich für die Forschung ist und Kausale Inferenz ebenfalls vielversprechende Werkzeuge mitbringt, hält der Autor es für wichtig, diese Werkzeuge so zugänglich wie möglich zu gestalten. Aus diesem Grund wurde im Rahmen dieser Arbeit viel Wert darauf gelegt, Prozesse transparent und nachvollziehbar zu beschreiben und - falls möglich - Entscheidungen verständlich zu begründen. Das Lesen dieser Arbeit soll den Grundstein für den Einstieg in das Thema Kausale Inferenz und Double Machine Learning, sowie das *DoubleML*-Package schaffen können, damit in Zukunft häufiger tatsächliche kausale Effekte statt Assoziationen untersucht werden. Der bereitgestellte Code kann durch den modularen Aufbau leicht abgewandelt und für andere Analysen verwendet werden. Wie im Hauptteil der Arbeit gesehen werden konnte, können Double Machine Learning und klassische statistische Methoden sich in ihren Ergebnissen unterscheiden. Aus diesem Grund wäre es zukünftig von großem Interesse, weiterhin Ergebnisse beider Ansätze auf identischen Datensätzen zu vergleichen. Ebenfalls kann es sinnvoll sein, ältere Studien von hoher akademischer Relevanz, falls möglich, unter Anwendung von Double Machine Learning erneut zu betrachten.

Anhang

Der vollständige Code zur Datenverarbeitung, Modellimplementierung und Analyse ist in der beigefügten Jupyter-Notebook-Datei auf dem elektronischen Datenträger enthalten.

Literatur

Acuna, E., & Rodriguez, C. (2004). The Treatment of Missing Values and Its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, 639–647.

Allison, P. D., et al. (2010). *Missing Data* (Bd. 200210). Sage Thousand Oaks, CA.

Angrist, J., & Imbens, G. (1995, Februar). *Identification and Estimation of Local Average Treatment Effects* (Techn. Ber.). National Bureau of Economic Research. https://doi.org/10.3386/t0118

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (n. d.). *DoubleML in Python*. https://docs.doubleml.org/stable/guide/basics.html (accessed: 01.03.2023).

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). DoubleML – an Object-Oriented Implementation of Double Machine Learning in R. https://doi.org/10.48550/ARXIV.2103.09603

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2022). DoubleML – an Object-Oriented Implementation of Double Machine Learning in Python. *Journal of Machine Learning Research*.

Baicker, K., Allen, H., Wright, B., Taubman, S., & Finkelstein, A. (2018). The Effect of Medicaid on Management of Depression: Evidence From the Oregon Health Insurance Experiment: the Effect of Medicaid on Management of Depression. *The Milbank Quarterly*, *96*, 29–56. https://doi.org/10.1111/1468-0009.12311

Baicker, K., & Finkelstein, A. (2014). *Oregon Health Insurance Experiment*. https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment?page=1%5C&perPage=50

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., & Finkelstein, A. N. (2013). The Oregon Experiment — Effects of Medicaid on Clinical Outcomes [PMID: 23635051]. *New England Journal of Medicine*, *368*(18), 1713–1722. https://doi.org/10.1056/NEJMsa1212321

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-parameter Optimization. *Journal of Machine Learning Research*, *13*(2).

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V., Stukel, T. A., Lucas, F. L., Batista, I., Welch, H. G., & Wennberg, D. E. (2002). Hospital Volume and Surgical Mortality in the United States. *New England Journal of Medicine*, *346*(15), 1128–1137. https://doi.org/10.1056/nejmsa012337

Bramer, M. (2007). Avoiding Overfitting of Decision Trees. *Principles of Data Mining*, 119–134.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bro, R., Kjeldahl, K., Smilde, A. K., & Kiers, H. (2008). Cross-validation of Component Models: a Critical Look At Current Methods. *Analytical and Bioanalytical Chemistry*, *390*, 1241–1251.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A Method for Assessing the Quality of a Randomized Control Trial. *Controlled Clinical Trials*, *2*(1), 31–49. https://doi.org/https://doi.org/10.1016/0197-2456(81)90056-8

Chambers, R. L. (2003, März). *Analysis of Survey Data*. Wiley. https://doi.org/10.1002/0470867205

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, *107*(5), 261–65. https://doi.org/10.1257/aer.p20171038

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2016, Juli). *Double/Debiased Machine Learning for Treatment and Causal Parameters* (Papers Nr. 1608.00060). arXiv.org. https://ideas.repec.org/p/arx/papers/1608.00060.html

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*. https://doi.org/10.2139/ssrn.1819486

Coase, R. (n. d.). Quote: Ïf You Torture the Data Long Enough, It Will Confess To Anything."

Cournapeau, D. (n. d.). *sklearn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed: 22.06.2022).

Croxton, F. E., & Cowden, D. J. (1939). *Applied General Statistics.* Prentice-Hall, Inc. https://doi.org/10.1037/13608-000

Davey Smith, G. (2009). Smoking and lung cancer: causality, Cornfield and an early observational meta-analysis. *International Journal of Epidemiology*, *38*(5), 1169–1171. https://doi.org/10.1093/ije/dyp317

Dickman, S. L., Himmelstein, D. U., & Woolhandler, S. (2017). Inequality and the Health-care System in the USA. *The Lancet*, *389*(10077), 1431–1441. https://doi.org/https://doi.org/10.1016/S0140-6736(17)30398-7

Durrett, R. (2019, April). *Probability* (5. Aufl.). Cambridge University Press.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Group, O. H. S. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year*. *The Quarterly Journal of Economics*, *127*(3), 1057–1106. https://doi.org/10.1093/qje/qjs020

Fradkin, D., & Madigan, D. (2003). Experiments with Random Projections for Machine Learning, 517–522. https://doi.org/10.1145/956750.956812

Georgii, H.-O. (2009, Juni). *Stochastik*. Walter De Gruyter. https://doi.org/10.1515/9783110215274

Granger, C. W. (1986). Statistics and Causal Inference By P. Holland. *Journal of the American Statistical Association*, *81*(396), 967–968.

Greenland, S., Pearl, J., & Robins, J. M. (1999a). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48.

Greenland, S., Pearl, J., & Robins, J. M. (1999b). Confounding and Collapsibility in Causal Inference. *Statistical Science*, *14*(1). https://doi.org/10.1214/ss/1009211805

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A Critique of the Cross-lagged Panel Model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hawkins, D. M. (2003). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Heckman, J. (1977, März). *Sample Selection Bias As a Specification Error (with an Application To the Estimation of Labor Supply Functions)* (Techn. Ber.). National Bureau of Economic Research. https://doi.org/10.3386/w0172

Hernan, M. A., & Robins, J. M. (2023, Mai). *Causal Inference*. CRC Press.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A Practical Guide To Support Vector Classification.

Hussey, P. S., Schneider, E. C., Rudin, R. S., Fox, D. S., Lai, J., & Pollack, C. E. (2014). Continuity and the Costs of Care for Chronic Disease. *JAMA Internal Medicine*, *174*(5), 742–748.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kantarjian, H., & Rajkumar, S. V. (2015). Why Are Cancer Drugs So Expensive in the United States, and What Are the Solutions? *Mayo Clinic Proceedings*, *90*(4), 500–504. https://doi.org/https://doi.org/10.1016/j.mayocp.2015.01.014

Kohavi, R., & Wolpert, D. (1997). Bias Plus Variance Decomposition for Zero-One Loss Functions.

Malthus, T. R. (1888). *An Essay on the Principle of Population: Or, a View of Its Past and Present Effects on Human Happiness*. Reeves & Turner.

Nichols, A. (2006). *Weak Instruments: an Overview and New Techniques* (North American Stata Users' Group Meetings 2006). Stata Users Group. https://EconPapers.repec.org/RePEc:boc:asug06:3

Ott, J., Ullrich, A., & Miller, A. (2009). The Importance of Early Symptom Recognition in the Context of Early Detection and Cancer Survival. *European Journal of Cancer*, *45*(16), 2743–2748. https://doi.org/https://doi.org/10.1016/j.ejca.2009.08.009

Oxford Dictionary of English. (2010, Januar). Oxford University Press. https://doi.org/10.1093/acref/9780199571123.001.0001

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference* (2nd). Cambridge University Press.

Pearl, J. (2009b). Causal Inference in Statistics: an Overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-ss057

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies for Random Forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What To Do When Data Are Missing in Group Randomized Controlled Trials. NCEE 2009-0049. *National Center for Education Evaluation and Regional Assistance*.

Resnik, D. B., & Vorhaus, D. B. (2006). Genetic Modification and Genetic Determinism. *Philosophy, Ethics, and Humanities in Medicine*, *1*, 1–11.

Roberts, P., Priest, H., & Traynor, M. (2006). Reliability and Validity in Research. *Nursing Standard*, *20*(44), 41–45. https://doi.org/10.7748/ns2006.07.20.44.41.c6560

Romeijn, J.-W. (2022). Philosophy of Statistics. In E. N. Zalta & U. Nodelman (Hrsg.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: the Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591. https://doi.org/10.2307/2287653

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter Tuning and Performance Assessment of Statistical

and Machine-learning Algorithms Using Spatial Data. *Ecological Modelling*, *406*, 109–120.

SGB V. (1989).

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association*, *52*(280), 548–551. https://doi.org/10.1080/01621459.1957.10501412

Summers, J., & Morrison, E. (2009). Principles of Healthcare Ethics. *Health Care Ethics. 2nd Ed. Sudbury: Jones and Bartlett Publishers*, 41–58.

Syrgkanis, V., & Zampetakis, M. (2020). Estimation and Inference with Trees and Forests in High Dimensions. https://doi.org/10.48550/ARXIV.2007.03210

Thompson, R. P., & Upshur, R. E. (2017, August). *Philosophy of Medicine*. Routledge. https://doi.org/10.4324/9781315159843

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Verfügbar 21. März 2023 unter http://www.jstor.org/stable/2346178

VanderWeele, T. J., & Shpitser, I. (2013). On the Definition of a Confounder. *The Annals of Statistics*, *41*(1). https://doi.org/10.1214/12-aos1058

Vladeck, B. (2003). Universal Health Insurance in the United States: Reflections on the Past, the Present, and the Future. *American Journal of Public Health*, *93*(1), 16–19. https://doi.org/10.2105/ajph.93.1.16

von Winterfeld, D., & Edwards, W. (1986, September). *Decision analysis and behavioral research*. Cambridge University Press.

Wang, L., & Alexander, C. (2016). Machine Learning in Big Data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*, 52–61. https://doi.org/10.33889/IJMEMS.2016.1.2-006

White, F. (2015). Primary Health Care and Public Health: Foundations of Universal Health Systems. *Medical Principles and Practice*, *24*(2), 103–116. https://doi.org/10.1159/000370197

Wong, J., Manderson, T., Abrahamowicz, M., Buckeridge, D., & Tamblyn, R. (2019). Can Hyperparameter Tuning Improve the Performance of a Super Learner?: a Case Study. *Epidemiology*, *30*, 1. https://doi.org/10.1097/EDE.0000000000001027

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Yarlett, D. (2002). Uncertainty in Causal and Counterfactual Inference.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/https://doi.org/10.1016/j.neucom.2017.01.026

Abkürzungsverzeichnis

bzw. Beziehungsweise

cet. par. ceteris paribus

d. h. das heißt

E. D. eigene Darstellung

etc. et cetera

et al. et alia

IIVM Interactive IV Model

IRM Interactive Regression Model

IV Instrumentalvariable

n. d. no date

OHIE Oregon Health Insurance Experiment

OLS Ordinary Least Squares

z.B. zum Beispiel

Symbolverzeichnis

ε Störterm

ω kausaler Effekt

ϱ Zufallsfehler

ϑ Standardabweichung

ϖ kausaler Effekt

Junior Management Science

www.jums.academy
ISSN: 2942-1861

The Dark Side of Employer Branding – Aesthetic Labour and Employer Attractiveness in the Beauty and Cosmetics Industry

Susanne Rautzenberg

Vienna University of Economics and Business

Abstract

This study explores the impact of Aesthetic Labour on employer branding within the cosmetics industry. Traditionally, beauty and cosmetics firms are associated with positive connotations such as elegance, glamour, and attractiveness, driven by professional marketing strategies. Aesthetic Labour, the employment of workers who visually represent the company's brand image, extends beyond customer-facing roles to influence the overall company image and employer brand. This thesis investigates how these Aesthetic Labour practices shape perceptions and expectations of prospective employees, particularly focusing on Generation Z, who prioritize diversity and inclusion.

Through semi-structured interviews with twelve Generation Z employees and former employees of well-known cosmetics firms, this research uncovers the dual nature of the industry's employer branding, one being the glamorized image that is carried to the outside, the other highlighting the downsides of the glamorized effect. While a polished, glamorous image attracts applicants, it often leads to unrealistic expectations about the work environment. The study finds that although employees experience a supportive and enthusiastic work culture, they also face internal pressure to maintain a specific aesthetic standard. This pressure, albeit not officially mandated, is perpetuated through internalized norms and possibly biases in the hiring process.

The findings reveal that Aesthetic Labour practices influence employer attractiveness by creating a desirable but often unattainable image. This study contributes to the understanding of Aesthetic Labour's role in shaping employer brand perceptions, highlighting the need for cosmetics firms to balance their branding strategies with inclusive and diverse employment practices to meet the evolving expectations of the modern workforce.

Keywords: aesthetic labour; employer branding; cosmetics industry; generation z

1 Introduction

Beauty and cosmetics firms come with a reputation, standing for elegance, glamour, graceful aging, attractiveness, aesthetics, and much more. It is hard to miss that all the words related to the word beauty and cosmetics are positively connotated, only rarely inducing negative feelings. This may be due to the professional marketing of not only the products themselves but also the marketing strategies of the companies behind the products, which for years have been created to perfection by professionals to seem desirable, appealing, and prestigious. For centuries, beauty standards have existed, and cosmetics firms' promises to help users obtain the idealized look have certainly added to the industry's glamorous feel. Although cosmetic products themselves have been around for thousands of years, the market capitalization of the beauty industry started to gain traction after the 1920s (Peiss, 2000). Marketing beauty standards to households became the standard program of every cosmetics firm, leading to a huge increase in consumption and driving the value of the beauty industry up to USD 430 Billion globally in the year 2022 (McKinsey & Company, 2023).

Although most marketing campaigns were aimed at attracting the consumer – mainly females – to the product itself, every bit of the brand's representation to the outside is displayed in product marketing campaigns. Especially at beauty counters, employees that represent the brand to the consumer adhere to certain beauty standards to be able to sell the product better, silently saying: "By buying our products, you could also look like this." This concept is known as *Aesthetic Labour (AL)*. This means that employees are hand-picked by the companies, depending on whether they have the right look according to company standards and helping to shape their brand image (Warhurst & Nickson, 2007a). The term was most prominently defined by Chris Warhurst and Dennis Nickson in 2000, originating from earlier research done on emotional labor by Arlie Hochschild, which examined how workers manage their emotions to meet job requirements. Aesthetic Labour extends this idea to the visual and physical attributes that employees bring to their roles, particularly in the service sector (Warhurst et al., 2000). It is plausible that

DOI: https://doi.org/10.5282/jums/v11i1pp139-163

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

this exact image may have expanded from frontline workers into the office, creating a company image with a similar perceived aesthetic. This concept influences not only the brand but also the company image, therefore also influencing the employer brand (EB).

The concept of the employer brand is derived from marketing literature and first became relevant in the 1990s when several marketing experts began to research the importance of retaining employees for a company's profit due to the war of talent during the late 20th century (Martin et al., 2011). More and more, employees were referred to as "internal customers," which is how the first concepts of the Employer Brand started being developed. In these concepts, employees are viewed as customers, explaining that "it is easier, cheaper, and more profitable to keep existing customers than recruit new ones" Ambler and Barrow, 1996, p. 185. Ambler and Barrow first defined the employer brand as the combination of "functional, economic and psychological benefits provided by employment" Ambler and Barrow, 1996, p. 187, which has laid a foundation for countless further studies that explore the way companies and jobs are viewed by (potential) employees Alnıaçık and Alnıaçık, 2012. These studies entail keywords like *Job Attractiveness, Employer Branding, Employer Image, Internal Marketing,* and *Job Perception* and have been significant in changing the way that Strategic Human Resource Management (SHRM) is viewed from a scientific perspective (Martin et al., 2011).

Even now, much of recent research still refers to these findings. Ambler and Barrow, 1996 found that the concept of "Employer Brand" is primarily seen as equivalent to corporate culture by employees. Key components for creating a good EB include creating awareness of the company itself, so the "company's performance, market position, reputation and product brand(s)" Ambler and Barrow, 1996, p. 197, as well as psychological benefits, which are derived from these aspects. Both aspects enable a company to "recruit, retain and motivate the best people" Ambler and Barrow, 1996, p. 197. Building from the employer brand concept, academics later used these findings to create models that showcase what (prospective) employees value in an organization. The most significant model that has been derived from Ambler and Barrows findings is the Employer Attractiveness model (EmpAt in short) by Berthon et al., 2005, including the dimensions of employer attractiveness, which are impacted by the employer image, which in turn is shaped by EB practices, as well as external aspects and internal realities (Barnett et al., 2006). The way that AL plays into the employer brand shall be researched in this thesis and may put into light a darker side of employer branding.

1.1 Problem Definition

Aesthetic Labour has long been a concept that is viewed critically in the academic world. Rating employees so superficially leaves a bad aftertaste, especially as a new generation, Generation Z (short: Gen Z), is entering the labor market, demanding employers to represent topics like diversity, inclusion, or sustainability (McKinsey & Company, 2023). Parallels to the beauty industry can be drawn through the fashion industry, where even in 2024, brands like *Brandy Melville* are getting exposed for not selling inclusive clothing and only employing a certain type of person – one that matches the company and brand image: *young*, *skinny*, *white* (Die Presse, 2024). The fact that this image is harmful and outdated is hardly negotiable. A similar discussion has been ongoing for years with the fashion giant *Abercrombie & Fitch*, which was first sued in 2004 for discriminating against people of color, allowing them to work only in the back storage rooms, again in 2009 and 2015, and most recently in late 2023 (Fonrouge, 2023). Every time, the lawsuits entailed discrimination against employees who did not have the typical look or abuse and sexualization of those who did. As the ex-CEO of the firm, Mike Jeffries, stated in 2006: *"We go after the cool kids. We go after the attractive all-American kid with a great attitude and a lot of friends. A lot of people don't belong [in our clothes], and they can't belong. Are we exclusionary? Absolutely"* (Mike Jeffries 2006 in Loftus 2013, 11. 5ff), openly excluding anyone not fitting his image of the company.

Most literature focuses on frontline employees of the retail and hospitality industries, mainly feeding from scandals that arise every few years. Aesthetic Labour literature has rarely touched on looks in an office setting, nor has the beauty and cosmetics industry been closely looked at. May this be due to a lack of scandals in the beauty industry? In the fashion industry, articles are regularly published, displaying the lack of inclusive sizes (Dion & Tachet, 2019), greenwashing (Alizadeh et al., 2024), or the specific exclusion of certain target groups (Lies, 2021); *Brandy Melville* and *Abercrombie & Fitch* being prime examples of the topic's relevance even today. More quietly, scandals formed in the cosmetics industry, including scandals concerning animal testing or the exclusion of certain groups. Over the years, however, it has become a norm for cosmetic products to be labeled as vegan or cruelty-free (Santos et al., 2023) and well-known beauty brands like *MAC Cosmetics* or *Fenty Beauty* have increased their product ranges to include over 40 skin shades each (Frisby, 2019), showing a shift in the industry towards diversity and inclusion. Furthermore, diversity and inclusion in branding and marketing campaigns have become much more relevant, highlighting the importance of an employer creating a space for everyone to feel included (Campbell et al., 2023). However, several authors argue that "pretty privilege," an expression that summarizes the privilege of conventionally attractive people, is still common, meaning that employees who fit the company's look and image are more likely to be hired (Talamas et al., 2016).

Especially in the cosmetics industry, one may quickly assume that the glamorous art of living that is portrayed through marketing campaigns depicts the way of life of that company's employees. Striving for a similar look, lifestyle, or aesthetic, applicants may look past aspects that are negatively connotated with the brand, or that would usually make an employer seem unattractive to them. The industry's general popularity cannot be denied, as firms like

the industry leader L'Oréal regularly publish statistics that showcase their high number of applications each year (e.g., L'Oréal 2020), whereas less glamorous industries, like the rail industry, struggle to find suiting talents, despite offering the same functional benefits (Wallace et al. 2012; see definition of functional benefits in chapter 2.1.1). That the cosmetic industry's companies are automatically connected to beauty and aesthetics is undeniable. Product marketing allows a consumer to see which looks they could attain through the product's use, creating a desire for a connection to that brand and a "sense of belonging" to a brand (Vonk, 2020). By having this demand for aesthetics on the one hand and a push for diversity and inclusion on the other, a new challenge for the industry is presented. AL as a framework can help in analyzing this problem.

1.2 Research Goal and Relevance

According to Warhurst and Nickson, 2007b, Aesthetic Labour is the "employment of workers with desired corporeal dispositions" (p. 107), ultimately contributing to the "production and portrayal of a distinct and defined corporate image" (p. 107). In combination with the earlier mentioned employer brand, the concept can be seen as a contributor to a desired perception and employer image to the outside. While most research in the field has been done on employees in the frontline service area, including jobs in hospitality and retail (Karlsson 2011), only limited research has been done on how the concept of Aesthetic Labour can be applied to office jobs and how frontline employees and photos that are broadcast through an employer website or marketing campaigns affect the employer image, in turn affecting the employer brand that is used to recruit new talents. Furthermore, there is barely significant literature focused on the beauty and cosmetics industry. EB is embedded in most companies to make them more attractive to applicants. Aesthetic Labour shall serve as a concept that will help in gaining a new perspective on EB, showcasing tensions and challenges that may arise from EB. The way this shall be done is through gaining an in-depth understanding of Aesthetic Labour, which mainly involves the appearance and demeanor employees are expected to maintain in certain industries as a part of their job roles. By understanding this concept and how it impacts brand image, this thesis will add depth to existing literature on Aesthetic Labour while also serving as a paper that critically illuminates EB.

Understanding the connection between the two concepts and critically highlighting some aspects that influence the employer image will not only feed into a research gap but will also allow employers or human resource specialists to gain insights into how their intentional (as well as unintentional) branding practices influence their brand image and therefore affect talent attraction, retention, and motivation, allowing them to improve recruitment and employee well-being strategies, especially as Gen Z is taking over the labor market. To best uncover the connection of these concepts in the cosmetics industry, semi-structured interviews with people who are currently or have in the past been employed in a well-known cosmetics firm shall be conducted. According to McKinsey & Company, 2023, new standards, including diversity and inclusion, have been defined by Gen Z, and this study will focus mainly on young talents. The specific focus on Gen Z employees will allow us to get an understanding of which image of the cosmetics and beauty industry is engrained in the minds of even the younger generation. Organizations will benefit in understanding what makes them attractive to young talents, as their "knowledge, skills and fresh insights drive innovation and, in the long run, competitive advantage" Ružić and Benazić, 2023, p. 97.

In this study, young talents from Gen Z are defined as people born no earlier than 1995, have less than four years of professional work experience, and have significant experience in the beauty and cosmetics industry. Within this core group, the EB concept, as well as aspects influencing the employer brand, will be researched. By combining EB with AL literature, it can be assumed that Aesthetic Labour practices, like marketing campaigns, employee representation, or media coverage, shape expectations in the type of employee as well as work inside the company, are valued when making the application decision and may form expectations, that are not fully met inside the company. These assumptions lead us to the following research questions:

> ***RQ1**: How do Aesthetic Labour practices in the beauty and cosmetics industry influence the perceptions and expectations of prospective employees?*

This research question aims to find out what applicants **expect** of a job in the industry concerning the type of person working inside the company, as well as the job inside the company itself. To gain information on both topics, interview questions shall lead to responses that entail perceptions of a company that were built before considering them as an employer, emotions that applicants had toward the brand, and first touchpoints with the company. A focus will be laid on brand image portrayed through Aesthetic Labour practices in marketing campaigns or through laborers. This first question aims to find out whether AL is expected by applicants before entering the firm as an employee.

Secondly, the **values** of employees will be evaluated to see whether they align with general findings from EB literature, as well as the earlier mentioned values like inclusion and diversity. Therefore, the second research question reads:

> ***RQ2**: To what extent do job seekers prioritize aesthetic aspects, such as brand image and lifestyle portrayal, when evaluating potential employers in the cosmetics industry?*

This second research question will help to understand how big the impact of aesthetic practices is on the employer brand. Interview questions will include comparisons and drawing parallels to other companies and job opportunities, as well as help to understand the influence of the aesthetics of the brand on the application decision-making process. Lastly, both the expectations and values shall be compared

to the reality, so the **experience** in the firm, once employed, leading to the last research question:

> ***RQ3****: How does the perception of a job in the cosmetics industry that is shaped by Aesthetic Labour practices differ from the experience within the company once employed?*

With this last question, we aim to find out how much the brand image of a prospective employee (before employment) and their expectations overlap with the experience inside a company. This will help in distinguishing whether Aesthetic Labour adds to employer branding in a positive, negative, or neutral way and how the concept influences the overall brand image. This question aims to figure out whether the perceived image outside the firm seems more glamorous than the job itself is and will help uncover discrepancies between brand image and reality.

As mentioned above, this thesis aims to investigate the elements that contribute to employer attractiveness in the cosmetics industry, focusing on the effects of employer perceptions that stem from Aesthetic Labour practices. By conducting qualitative research in this area, employees of cosmetics firms have an opportunity to add their perceptions and experiences, creating a holistic picture of a cosmetic company's image from the application to the job itself. As the connection of Aesthetic Labour to employer image, specifically for office jobs, has never been investigated, this research will fill a gap in existing literature. Additionally, practical insights for cosmetics companies looking to enhance their employer branding strategies to attract and retain young talents can be provided.

1.3 Structure of the Thesis

The way this thesis shall be structured to finally be able to answer the above-named research questions is the following. First, a literature overview of all earlier mentioned concepts will be conducted. At first, employer branding, the formation of the employer image, and factors of employer attractiveness will be explained to gain a holistic understanding of the way employers themselves can actively shape their organizational image. After that, Aesthetic Labour as a framework and the ideal self will be investigated to critically highlight EB. A general understanding of the strategic portrayal of a certain lifestyle or persona and how that affects a (prospective) employee's attraction to one firm will also be provided by this. Finally, the connection of both concepts will be laid out to understand how they impact one another. This literature overview will help in understanding the impact of Aesthetic Labour on the employer image. Following the literature overview, the research objective will be formulated. After that, the methodology for answering the before-defined research questions, including the research design and process, data analysis method, as well as critical factors and limitations in the process, will be laid out. After that, the findings of the conducted research will be described to discover potential patterns, themes, or categories in the collected data. In the following section, there will be a discussion including the interpretation of these findings, and from those, practical implications and directions for future research will be identified. A short conclusion will round up this thesis.

2 Theoretical framework

To be able to gain a broader understanding of the research that has been done in all of the above mentioned areas, a thorough literature review will be done in this chapter. All literature has been peer-reviewed and published in serious academic journals, including but not exclusive to journals like the *Journal of Brand Management*, the *International Journal Of Advertising, Personnel Review,* the *Journal of Marketing*, or the *International Journal of Human Resource Management.* Several of these journals focus on marketing-related topics, as employer branding actually stems from marketing literature rather than human resource management (HRM). However, the concept's relevance in HRM will be looked into further in the following chapters.

Literature was identified mainly through platforms like Google Scholar, which led to academic articles across the research platforms ELSEVIER, Springer, Wiley, SAGE Publishing, IEEE Xplore, ScienceDirect, and more. Articles were filtered by year, first excluding any literature that appeared before 2010. It quickly became clear that many of the still-used and relevant models were created and discussed before this date, which is why older articles and research were also taken into account. However, recent studies using these models were always included in the search. Going from there, the snowballing method was used to find more suitable research on all topics.

To begin this chapter, the concept and backgrounds of employer branding from the employer's point of view will be looked into in detail. This includes the general concept of employer branding and its use in recent research. As employer branding impacts the *image* of a company as well as its *attractiveness* as an employer, both of these concepts were looked into in more detail. As the aim of this thesis is to research the impact of the Aesthetic Labour concept on the employer brand, the concept of *Aesthetic Labour* was investigated thoroughly. To explain its connection to employer attractiveness, the *ideal self* was also researched.

Keywords that were used to find suitable research on these topics include 'employer brand,' 'employer branding,' 'employer attractiveness,' 'employer image,' 'employer reputation,' 'Aesthetic Labour,' and 'ideal self'.

2.1 Employer Branding in Companies

The concept of the employer brand originally stems from general branding practices, which are a part of marketing literature. Applying branding practices to HRM is called *employer branding* in literature (Backhaus & Tikoo, 2004). To gain a broad overview of Employer Branding (short: EB), an introduction including a definition, relevant studies, and related concepts will be given in the following.

2.1.1 Employer Branding

Employer branding and the employer brand first became a term in the early 1990's. Defined for the first time by Ambler and Barrow in 1996, the employer brand is "the package of functional, economic and psychological benefits provided by employment, and identified with the employing company" Ambler and Barrow, 1996, p. 187. The reason for their extended research on employer branding practices was the realization of the correlation between the good quality of employees and the quality of the sold product or service. The researchers divide the employer brand into three main factors that a brand can offer a potential employee and that are parallel to the benefits "that a conventional (product) brand offers to consumers," namely functional (what is the use), economic (material or monetary rewards), and psychological (feelings like belonging, direction, or purpose) benefits Ambler and Barrow, 1996, p. 186f. For a firm to attract as many good potential employees as possible, their employer brand should cover all three of these benefits. The researchers base their findings on concepts similar to EB: corporate culture and identity, internal marketing, and corporate reputation. Within their research, they applied the concept of EB to each of these (Ambler & Barrow, 1996).

Corporate *culture and identity* are defined by Hiller and Verdier, 2014 as the "shared beliefs and preferences among members of one organization," particularly between different levels of hierarchy, creating a strong corporate identity for individual employees Hiller and Verdier, 2014, p. 94. *Internal marketing* is *what* the company does to create a stronger corporate identity for their employees, or as defined by Rafiq and Ahmed, 2000, "a planned effort using a marketing-like approach to overcome organizational resistance to change and to align, motivate and interfunctionally co-ordinate and integrate employees towards the effective implementation of corporate and functional strategies to deliver customer satisfaction through a process of creating motivated and customer orientated employees" Rafiq and Ahmed, 2000, p. 454. Lastly, *corporate reputation* is the way a company is viewed not only internally by its employees but also externally by stakeholders, which may impact the way the general public and (potential) employees view the company (positively and negatively). According to Cruise O'Brien, "reputation is developed based on the observation and assessment of consistent behavior over time" (O'Brien 1995 in Ambler and Barrow 1996, p. 190), indicating that the more positive attention a company attracts, the more positive its reputation will be, making it a more attractive employer.

It is clear as to why Ambler and Barrow decided to disclaim the possible connection of these concepts to that of the employer brand. Although all three concepts are different from one another, they have certain overlaps that the EB concept can combine. In EB, the "similarities between HR and regular (product) brand marketing" are recognized, therefore permitting "their functional skills to be used in each other's areas" Ambler and Barrow, 1996, p. 190. The concept is needed to keep employee commitment high for existing employees as well as attract new ones.

Ambler and Barrow's definition of EB was later used as a foundation of numerous studies in the field and continuously developed by further researchers. Sullivan re-defined EB in 2004, calling the concept "a targeted, long-term strategy to manage the awareness and perceptions of employees, potential employees, and related stakeholders with regards to a particular firm" (Sullivan 2004 in Backhaus and Tikoo 2004, p. 501), which depicts a more accurate definition of what EB actually does: manage awareness and perceptions. In the same year, Backhaus and Tikoo defined employer branding as a process of "building an identifiable and unique employer identity, and the employer brand as a concept of the firm that differentiates it from its competitors" Backhaus and Tikoo, 2004, p. 502 by applying general branding principles to HRM. Later, Edwards, 2010 explained the term by comparing it with principles from marketing and branding that are applied to HR initiatives (Edwards, 2010), emphasizing Ambler and Barrow's internal marketing approach. This is further supported by Ružić and Benazić, 2023 research on employer attractiveness – they state that a marketing approach is the most effective in attracting people by effectively implementing marketing initiatives to obtain a wanted behavior from their target audience (Ružić & Benazić, 2023); in this case, the job application to the company. Sivertzen et al., 2013 later defined EB as the process of "building employer identity directed at existing and potential employees, to differentiate the firm from its competitors" Sivertzen et al., 2013, p. 474. All in all, it can be said that the concept of EB is a process that applies marketing and branding principles to HR practices to manage the perception of a company, both internally to existing employees and stakeholders, as well as externally to potential new employees. To break down both sides a bit more and make the understanding of the concept easier, we will mainly focus on the external audience in this thesis, where EB "encompasses how the organisation exposes itself and how others evaluate the organisation as an employer" and how "employer branding is used to increase employer attractiveness and improve corporate reputation" Sivertzen et al., 2013, p. 474. Generally, EB is "about promoting a clear view of what makes an organization desirable and different from an[other] employer" (Theurer et al., 2021).

The way EB is normally implemented into an organization is by applying branding principles that stem from marketing to HRM. Backhaus and Tikoo, 2004 describe this as a three-step process – first, a value proposition is developed, stating the firms' most attractive values, taking into account things like employer image, organizational culture, and impressions of product or service quality. Later, this value proposition is marketed to potential employees or recruiting agencies. In the last step, the value proposition is also marketed internally to ultimately create a workforce that is motivated by the company's values and hard to imitate by others. The goal of this is not only to attract new potential employees to a firm but also to encourage existing ones to stay with and support the company. Furthermore, Backhaus and Tikoo argue that marketing an employer

brand not only targets the wanted talents but also enhances the product and corporate brands (Backhaus & Tikoo, 2004). Presumably, this is a mutual relationship, indicating that the quality and popularity of the product's brand also affect the popularity of the organization as an employer. Generally, the two authors summarize that "employer brand associations form **brand images**, which in turn, affect **attraction** to the firm" Backhaus and Tikoo, 2004, p. 507, two concepts that will be looked into in more depth in the following.

To conclude, employer branding is the practice of curating an image of a company as an employer to make it more attractive to potential suitable applicants. To do so, general branding and marketing approaches are applied to HRM to actively shape the image of the company as an employer.

2.1.2 Employer Image

By managing awareness and perceptions as an employer, a certain image of the brand is created. The term *employer image* is often used synonymously with the term *employer reputation*. However, several authors argue that there is a small difference between the two. Barnett et al., 2006 note that the "image is the perception that external observers have of the firm," whereas "reputation is the net result of the aggregation of these perceptions" Barnett et al., 2006, p. 28. Sivertzen and colleagues simply describe employer reputation as a valuable resource in the resource-based view, representing a set of characteristics that have been projected onto an organization by the general public due to previous actions, ultimately leading to a competitive advantage or disadvantage (Sivertzen et al., 2013). As this definition does not specify whether this projection of characteristics occurs at a single point in time or over a longer period, it aligns more closely with Barnett's definition of employer image. This realization highlights the difficulty in distinguishing between both terms.

According to Barnett et al., "image can be shaped but not controlled by an organization" Barnett et al., 2006, p. 34 due to external factors that influence the company's reputation. A company can implement branding practices to improve the employer image, essentially painting a picture of the organization and communicating it not only internally but also externally (Sivertzen et al., 2013). Creating a coherent organizational image will affect not only employee attraction to a company but even more customer and consumer attraction. The more aware a consumer is of a brand and the more likely they are to use and identify with that brand's products, the more likely they are to later apply for a job at that brand, so shaping an employer image really begins by shaping a brand image (Collins & Stevens, 2002).

Collins and Stevens, 2002 also distinguish two dimensions of employer brand image, one being the general attitudes toward a company and the other being perceived job attributes by prospective employees, that are shaped by publicity like paid media, sponsorships, positive word-of-mouth endorsements and recruitment advertising (Collins & Stevens, 2002). Lievens and Slaughter, 2016 defined organizational image as an outsider's beliefs, whereas organizational insiders', so mainly employees', company image was compared with *organizational identity* (Lievens & Slaughter, 2016). Generally, it can be said that employer image is shaped by a combination of external perceptions and internal realities, reflecting how outsiders view the organization based on visible and communicated aspects. In the course of this thesis, the terms employer, organizational, and brand image will be used interchangeably, as they all contribute to shaping one image.

Researchers have long been studying employer image, trying to find out what the factors impacting the image are and how they can be internally controlled. A summary of diverse findings tells us that a company's image should be consistent over time to increase credibility (Theurer et al., 2021), attract and retain talents (Biswas & Suar, 2016) whilst also being personal enough for employees to identify themselves with the image, internalizing it and living for it more strongly (Brannan et al. 2015). Maintaining a positive image generally helps an organization for recruitment and retention purposes of high-performing employees, enhancing employee performance and building organizational prestige for a competitive advantage, in turn attracting even better talents, improving customer perceptions and investor confidence (Biswas & Suar, 2016). Aspects that positively impact the employer image include the level of innovativeness (Backhaus & Tikoo, 2004), corporate social responsibility (CSR) initiatives (Lievens & Slaughter, 2016), and effective communication of brand values (Sivertzen et al., 2013).

In studies, Sivertzen and colleagues (2013) describe the employer image from a resource-based view, indicating that the image is an important factor in contributing to a competitive advantage, highlighting the strategic importance of managing the employer image. In another study, Brannan and colleagues (2015) explored how the employer image affects identity work. The researchers focused specifically on the service sector, in which employer branding encourages employees to internalize the desired brand image, influencing both their self-identity and organizational identity. This study specifically highlights the *ideal self*, which is a version of the self that one wishes to be. By being able to identify with an organization's brand image, employees become more motivated to work there as they are more committed to organizational goals (Brannan et al. 2015). Biswas and Suar, 2016 also took a social identity approach, where employer image is especially impacted by organizational prestige, which stems from diverse information sources. These may include "reference groups' opinions, words-of-mouth, public relations, external controlled information, and internal communication regarding the outsiders' perception about the company" Biswas and Suar, 2016, p. 60. The more a (potential) employee identifies with this information, the higher the organizational prestige and, therefore, the work motivation as well as pride to be employed at the said firm becomes (Biswas & Suar, 2016). In this research, it becomes especially clear how important the alignment of external perceptions with internal realities is to maintain a positive employer image.

Especially in the hospitality and retail sectors, where Aesthetic Labour practices are common, workers are employed to help shape a company's image. This way, "image-oriented service managers [...] shape the brand's value proposition and differentiate themselves from competitors" Pounders et al., 2014, p. 670. Even one of their first academic papers on the concept of AL by Warhurst and Nickson in 2007 emphasized how important Aesthetic Labourers were for the "production and portrayal of a distinct and defined corporate image" Warhurst and Nickson, 2007b, p. 107, showing the impact of the concept on company perception to the outside.

In conclusion, the employer image is dynamic and influenced by various external aspects. Effective management of the employer image is integral to employer branding, contributing to the attractiveness of the organization as an employer. Reis et al., 2017 demonstrated that employer attractiveness is formed through a combination of instrumental aspects, such as job security and work-life-balance, and symbolic elements, which are "characterized as human traits that applicants attribute to the employer, such as sincerity, cheerfulness, and excitement" Reis et al., 2017, p. 1965. Studies like these show the intertwinement of the employer image with employer attractiveness, which will be discussed more in-depth in the following.

2.1.3 Employer Attractiveness

To be able to distinguish one company from various competitors and appear as the most attractive employer, companies can use the employer attractiveness model when curating an image that they would like to carry to the outside in their EB practices, shaping the employer image. Essentially, EB is done by a company to portray an image to applicants and (prospective) employees, which directly influences employer attractiveness (Reis et al. 2016). Soon after EB became relevant in the late 90s and early 2000s, researchers started focusing on elements that make an employer attractive to employees. One of the first models to effectively summarize the most important elements of employer attractiveness was developed by Berthon, Ewing, and Hah, who in 2005 developed the employer attractiveness model, or *EmpAt,* in short. Their study laid a foundation for organizations to adapt their EB practices and shape their brand image to be able to attract the "right talents" even more effectively and gain a competitive advantage in the recruitment of skilled employees (Lassleben & Hofmann, 2023). Berthon and colleagues defined employer attractiveness as "the envisioned benefits that a potential employee sees in working for a specific organization" Berthon et al., 2005, p. 156 and conducted a study on over 680 university students to create a five-factor model. Originally, 32 items were identified and then clustered into groups and ranked by relevance. The groups are *interest value*, *social value*, *economic value*, *development value,* and *application value*. Interest value describes a relatively subjective value, with its most relevant item being the employer's valuing of the employee's creativity. Social value describes the relationship that one employee may have with other colleagues, the most important items being a good relationship with colleagues and having supportive or encouraging colleagues. Economic value best describes the value of remuneration, compensation packages, and more. Development value describes the possibilities for personal and professional growth and development of an individual, most importantly, allowing the employees to feel good about themselves. Lastly, the application value depicts how "hands-on" an employee can work and apply newly learned skills within the organization (Berthon et al., 2005). In a literature overview done on employer attractiveness by Anke Dassler et al. in 2022, attractiveness factors were summarized and again divided into three groups: *transactional attributes*, including compensation, benefits, training, and development; *relational attributes*, including leadership style, work-life balance; and lastly *ideological attributes*, much like symbolic elements as described by Reis et al. These include the company brand, ethics and CSR, social media presence, and reputation (Dassler et al., 2022).

As a special focus in this thesis lies on young talents, special attention was laid on more recent studies that research employer attractiveness to young professionals, especially from Generation Z. Gen Z is defined as anyone "born after 1995, are in their early 20s and have just finished, or are about to finish, their university education" Lassleben and Hofmann, 2023, p. 546. According to a study done by Lassleben and Hofmann, 2023, Gen Z values most a fun work environment, a positive team environment, supportive relationships with colleagues and superiors, as well as flexibility concerning work hours and locations, attractive remuneration and investments in their development (Lassleben & Hofmann, 2023). Ružić and Benazić, 2023 confirm in their study that "acceptance and a good relationship with colleagues" (p. 95) is the most important factor for Gen Z in the workplace. The two researchers discovered that material working conditions like salary are less important. It is interesting to see that this study found that the factor that is least important to Gen Z is an organization's market activities, including innovative and high-quality products or services (Ružić & Benazić, 2023), an aspect that was previously named as positively impacting employer image (see chapter 2.1.2 Employer Image). Alnıaçık & Alnıaçık had similar findings in 2012 – in their study, they researched the different levels of importance of several factors of unemployed people vs. currently employed people. They, too, found that market value, including things like innovative and high-quality products or services or the opportunity to apply what was previously learned, was less important to Gen Z. They did, however, find out that the importance of this factor grew, the older the respondent was (Alnıaçık & Alnıaçık, 2012). Furthermore, Lassleben and Hofmann, 2023 found that Gen Z ranks aspects like opportunities to apply what has been learned, CSR, and other perks lower. However, as mentioned in the introduction, market research institutes have found the rising demand for diversity, inclusion, and sustainability, so CSR, in Gen Z (McKinsey & Company, 2023), indicating the need for further research on this generation. All in all, Gen Z values most the work environment, including their relationship with colleagues. Ac-

cording to several studies, market activities that shape the employer image are not as important.

To summarize, employer branding is what an employer actively does to positively impact their image and make them attractive as an employer. The employer image can be shaped by EB. However, it is additionally impacted by external factors that are not always in the hands of the organization. Employer attractiveness is shaped by the expectations a (prospective) employee has of an organization. The greater the overlap between these expectations and what is branded by an employer, combined with the company's image, the more likely a potential employee is to apply to that organization. Figure 1 helps in better showcasing the links between these three factors.

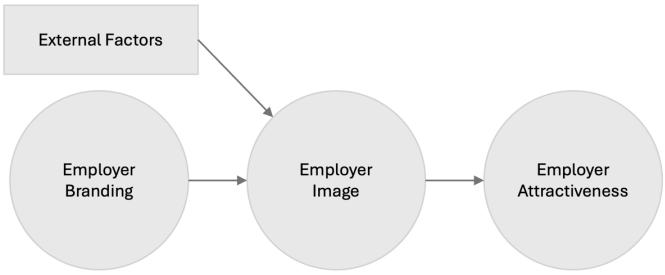

Figure 1: Employer Branding, Image and Attractiveness. Own Illustration.

One aspect that is rarely discussed in these studies is the impact of frontline workers on the brand as an employer and employer attractiveness. Oftentimes, brands decide to market their products by employing frontline workers who have the "right" look. This is especially done by retail clothing companies. Prominent examples of companies that implement this concept include brands like *Abercrombie & Fitch* or *Brandy Melville*, but can also be found in airline companies, hospitality, gastronomy, and more (Warhurst et al., 2000). This concept is also called *Aesthetic Labour* – the concept and its impact on EB will be highlighted in the following.

2.2 The Impact of Aesthetic Labour on Employer Branding

Aesthetic Labour was first defined by Warhurst, Nickson, and colleagues in 2000 when they started noticing a rise in job advertisements that required a certain look from the job candidate. This was predominantly noticeable in the retail and hospitality sectors but was later extended by researchers into the performing arts, hairdressing, or airline industries (Butler, 2014). Known companies that were often associated with these specific types of lookism include Disney, Monarch Airlines, Burton's Retail Group, Hooters, and others (Warhurst et al., 2000). The authors first described the concept as "a supply of embodied capacities and attributes possessed by workers at the point of entry into employment" Warhurst et al., 2000, p. 4. It is noticeable that all research done since this first definition by Warhurst and colleagues continues to focus on frontline workers in the service and hospitality sectors. So far, no research has been done on the impact of Aesthetic Labour on employer branding and its impact on organizational attractiveness. This is why, in the following, besides explaining Aesthetic Labour and its theoretical implications alone, the concepts of the *ideal self* will be explained to connect the concept of Aesthetic Labour to employer branding.

2.2.1 Aesthetic Labour

The concept of Aesthetic Labour has appeared in academic papers occasionally over the years. Each research on the topic will give a similar definition of the concept. While Mears defines Aesthetic Labour as "the practice of hiring employees based on their looks, which scholars have documented as a growing practice in the interactive service sector" Mears, 2014, p. 1330, the scholars Luoh and Tsaur, 2023 more recently defined it as "utilizing one's appearance or social attributes to create and present a professional and marketable image" Luoh and Tsaur, 2023, p. 1266. Both definitions emphasize looks being a part of a job that needs to be fulfilled. The researchers Entwistle and Wissinger, 2006 added personality as a part of Aesthetic Labour, impacting the way people are perceived as well. They conclude that a combination of both an aesthetic appearance as well as the right personality is necessary in Aesthetic Labour (Entwistle & Wissinger, 2006). Warhurst et al., 2000 also added to their original definition of Aesthetic Labour, implying that employers themselves shape and train aesthetic capacities and attributes into 'competencies' or 'skills,' resulting in a coherent 'style' of employee (Warhurst et al., 2000). Entwistle and Wissinger, 2006 criticize this, as personality as well as maintaining an aesthetic appearance take place outside of work, increasing pressure on employees. They also note that this pressure is more predominant in females rather than in males.

In general, researchers emphasize the value of soft skills over an employee's experience or technical skills, mainly focusing on subjective factors like perceived attractiveness, dress sense and style, as well as grooming, but also an employee's ability to relate well with people (Mears, 2014). Generally, most studies that have been done on Aesthetic Labour focus on the fashion, retail, and hospitality industries, as these are the ones with the most amount of customer contact through front-line workers, who are often the employees that are subject to Aesthetic Labour. In these studies, it is emphasized that Aesthetic Labour is not only about looking good but also about sounding right to attract the desired target group into shops or hotels (Warhurst & Nickson, 2007a). This subjective screening is already executed during the recruitment process; it can, however, also expand into the actual job through training, controls, or even sanctions, all part of a managerial strategy to shape the right brand image (Mears, 2014), promote products or services of organizations and sell an organizational image to customers (Entwistle & Wissinger, 2006). The more that the brand image that is shaped by employees aligns with customers' interests, the more likely these are to lower in-

terpersonal barriers and increase feelings like trust towards employees and the brand itself (Pounders et al., 2014).

The researchers Cutcher and Achtel (2012) researched exactly this problematic of the concept, citing recruitment pages like the Australian brand General Pants, stating that "ugly need not apply" (Swan 2012 in Cutcher and Achtel 2017, p. 675). As it is clear, this company implemented Aesthetic Labour for its brand to have more of an appeal for the "right" type of clientele to enter the store. With this approach, they cause "potential applicants to self-discriminate and de-select from jobs requiring an aesthetic they believe they do not possess" Cutcher and Achtel, 2017, p. 675f, a process which can be applied to many brands, whether they purposefully do so or unintendedly create an image like this. Cutcher & Achtel go on to explain that fashion retailers not only sell a product but more so a lifestyle, a statement that can be applied to the cosmetics sector, too. This includes hiring employees that "think regularly about the brand, about how to translate the brand and how to present themselves as artefacts of the brand" (Harquail 2006, p. 168 in Cutcher and Achtel 2017, p. 676). The researchers later carried out a study to gain more insights into the brand attachment of Aesthetic Labourers, involving 19 semi-structured interviews. They found that all respondents were positively tuned towards the brand when first recruited, identifying themselves with the brand's looks and values. Only later did they become less content with the brand itself after experiencing Aesthetic Labour as obligatory, having to adhere to rules and strict dress codes rather than allowing their style to be a simple part of their identity (Cutcher & Achtel, 2017).

The reason why Aesthetic Labour is particularly popular amongst employers is because it aids in curating a coherent brand image and attracts the targeted consumer. Furthermore, it has been proven that aesthetic elements like stylistic choices may increase perceived status and, therefore, power behavior (Baldessarelli et al., 2022). In a study done by Laura Vonk in 2020, the impact that employees have on the brand image was researched even further. To expand the findings by Cutcher and Achtel, 2017, where aesthetics, specific looks, and appearance are predetermined, Vonk analyzes how peers and colleagues influence aesthetics in the workplace. She does not limit her study to frontline and service sector employees but rather describes the work environment apart from the type of job. Instead, the author's focus lies on informal norms, organizational cultures, and peers that provide valuable insights into an organization's brand. Vonk also realizes that an aesthetic evaluation begins from the recruitment process onwards, where the selection of new employees is often based on a "cultural match," which also includes subjective factors like speech and style (Vonk, 2020).

Aesthetics, in general, significantly influence the market success of products and services. These aesthetic values, as noted by Lies, 2021, offer an added value beyond the functional benefits of a product, which is also known as *hedonic consumption*. This approach to aesthetics has many parallels with the creation of identity, established through affiliations such as "group membership, social class, and other demographic characteristics" Lies, 2021, p. 107. The aesthetic evaluation of a product often leads to a halo effect, where the positive perception of a product or a brand extends to enhance the brand and organizational image. Potential employees will mirror this perception of beauty with their "ideal selves" – the person they want to be and think they will be when using this brand's products and being employed within this organization (Lies, 2021).

In conclusion, Aesthetic Labour is the practice of hiring employees based on their looks and personalities to help shape and maintain a uniform company image. The implementation of this practice can lead to self-discrimination and de-selection from jobs requiring certain aesthetics (Lies, 2021), increased pressure to maintain aesthetic standards (Cutcher & Achtel, 2017), and increased power distance through colleagues (Vonk, 2020) in (potential) employees, showcasing the problematic of the concept. Why companies address the ideal self, nonetheless, will be discussed more in the following.

2.2.2 The Ideal Self

The concept of the ideal self helps in understanding not only consumer but also employee behavior. The following chapter will explore how a brand image will bridge the gap between the actual self (who people are) and the ideal self, who people aspire to be, through employer branding impacted by Aesthetic Labour.

According to Kim and Sullivan, 2019, the ideal self is often applied in a marketing strategy called *emotional branding,* which is a strategy used to better address customers and their affective states. This approach helps brands to build a strong consumer-brand attachment through sensory experiences or storytelling. For example, sensory branding leverages the sense of sight to create memorable experiences, while storytelling will involve consumers emotionally, allowing them to see themselves as a part of a brand's narrative. Experiences like these will help consumers better identify with a brand, as people are always looking for products and experiences that resonate with an image of their "ideal self" (Kim & Sullivan, 2019). Furthermore, emotional experiences help create a brand personality, which can also be used to convey the ideal self – the brand personality may be mirrored in an individual's personality, creating a higher bond and loyalty to one brand (Guthrie et al., 2008). The aesthetics of products can also be linked with the ideal self. Lies, 2021 argues that choosing products with an appealing design affirms the consumer's sense of self, enhancing their self-esteem and aligning with their ideal self. Lies names fashion companies that exploit this by marketing their products as tools to enhance attractiveness, thereby helping consumers achieve their aspirational selves (Lies, 2021). This implies that the concept of the ideal self can also be applied to the cosmetics sector, which is known for selling products to enhance beauty. Not only will the use of products help achieve the look of the ideal self, but it will also aid in mirroring either

their own personality or one they aim to achieve (Guthrie et al., 2008).

Cutcher and Achtel, 2017 furthermore discuss how the ideal self is not only prominent in consumers themselves but also in the workplace. They focus on Aesthetic Labourers in the retail sector who strive to live up to the ideal self the brand represents, modeling clothes and adopting brand standards as a part of their personal and professional identities. This work allowed them to develop a certain taste in clothing, a sense of fashion, and a personal aesthetic, allowing them to become subject to "the admiring gaze of others," giving them the role they were "striving to live up to... the self [they] would like to become" Cutcher and Achtel, 2017, p. 687. However, some of these Aesthetic Labourers did voice fatigue beginning to form, as they sometimes felt they were playing a role of their ideal self rather than portraying their authentic selves. Vonk, 2020 extends this discussion onto the organizational level, illustrating how aesthetic norms and cultural fits within a company influence how well an employee will fit in or how well they are perceived and valued within the company. Aesthetic norms that belong to the evaluation of a potential employee are formed by peers and can often showcase status inequalities, highlighting the complex dynamics of Aesthetic Labour, the ideal self, and the cultural norm of certain companies (Vonk, 2020).

Malär et al., 2011 named the example of the cosmetics industry and their branding efforts on consumers, stating that cosmetics have long been marketed as products that enhance beauty and bring consumers closer to an ideal version of themselves. Notably, the authors also introduce the term "actual self," which has been becoming more prominent in recent years. This concept focuses more on the portrayal of an imperfect version of a target group, as consumers are looking for more authenticity in marketing, something especially the brand *Dove* has been implementing for years. After conducting their study, the authors did find higher levels of emotional brand attachment when branding efforts focused on the actual self. Still, aspirational branding to achieve the ideal self works well, however, mainly on consumers with low involvement, self-esteem, or self-consciousness (Malär et al., 2011).

The ideal self is a concept that explains people's need to fit into their environment and become the version of themselves they strive to be. This chapter has explored how emotional branding, product aesthetics, and workplace norms, all creating a version of a brand that portrays a desirable lifestyle, leverage the ideal self to create strong consumer and employee engagement. However, some studies critically evaluate the concept, scrutinizing it for being outdated and pointing out that consumers build up a much higher level of emotional attachment when campaigns focus on the actual self. By understanding and addressing the aspirations of the ideal self, as well as its shift towards the preference of the actual self, companies can better align their branding strategies with the desires and motivations of their audience. This integration of the ideal self into branding and organizational culture is intricately connected to Aesthetic Labour, demonstrating how brands and workplaces shape and are shaped by the aspirations of individuals. It can be concluded that the portrayal of an ideal persona will shape not only the employer's image but also the expectations of a (potential) employee.

2.3 Conceptualizing the Links between Aesthetic Labour and Employer Branding

The above-named concepts can all be connected. While employer branding, image, and attractiveness are very closely related in a construct where one influences the other, the concepts of Aesthetic Labour and the ideal self mainly impact employer image. In Figure 2, I show the intertwinement and reciprocal effects of the following concepts: Employer Branding, Employer Image, Employer Attractiveness, Aesthetic Labour, as well as the Ideal Self.

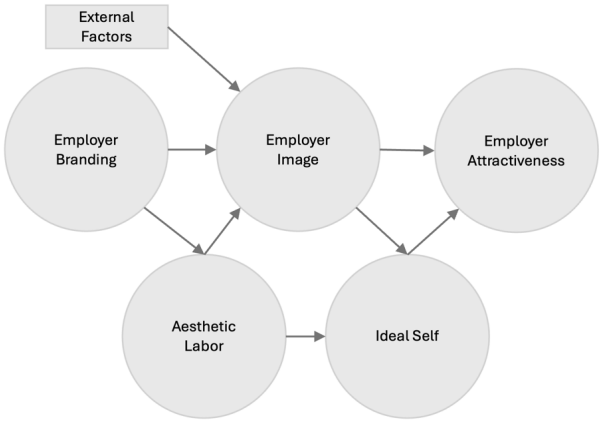

Figure 2: Connection of Employer Branding and Aesthetic Labour. Own Illustration.

Employer branding integrates principles of marketing to enhance a company's image as an attractive place to work. Aesthetic Labour plays a significant role in this context by intentionally contributing to the formation of an employer image that promises psychological benefits like a sense of belonging or purpose next to functional and economic benefits (Ambler & Barrow, 1996). The aesthetic appeal of employees can enhance a company's image by embodying the qualities the brand portrays. This specifically applies to sectors like fashion retail and hospitality, as well as beauty retailers, where employees' appearances can reflect a brand's prestige and aesthetic values, thus attracting customers who aspire to similar ideals (Cutcher & Achtel, 2017). By creating a self-identity for employees through the application of AL, employees will have an easier time imagining themselves as a version of their ideal selves in that organization. This self-identity will enable "a sense of security and a source of meaning" (Brannan et al. 2015, p. 32), rendering an employer more attractive. As "workers are likely to seek opportunities and relations that reinforce their sense of self, rather than those which contravene these notions" (Brannan et al. 2015, p. 33), the importance of focusing on the

portrayal of a certain lifestyle in employer branding is underlined. Through good EB, employees can identify with a brand in two ways: their *social self* as a professional skilled worker and their *ideal self* (Brannan et al. 2015). Furthermore, the Aesthetic Labour performed by employees helps in maintaining a consistent brand image that consumers recognize and trust. The alignment of employee appearance with brand values can reinforce consumer perceptions of quality and desirability, directly influencing brand image (Pounders et al., 2014).

To summarize, AL acts as a concept influencing and being a part of the EB strategy. Primarily, this strategy mainly targets consumers of the brand. However, exactly these consumers are the ones creating a company- and, therefore, employer image in their heads. To what extent exactly the employer image in the beauty industry is impacted by this concept shall be investigated further in the following study.

3 Research Objective

The primary objective of this thesis is to investigate whether aspects of Aesthetic Labour exist within the cosmetics industry and how this impacts the employer image and expectations towards companies in the industry as an employer. This investigation aims to bridge the gap between employment expectations that may be formed through the implementation of AL and the reality inside the company. The concept of Aesthetic Labour was chosen to highlight employer branding in the beauty and cosmetics industry from a critical point of view.

Through the preceding literature overview, assumptions can be made that connect EB to AL. First, it can be assumed that Aesthetic Labour practices, like marketing campaigns, employee representation, or media coverage, shape expectations in the type of employee and work inside the company. To see whether this assumption is true, research question number 1 is posed:

> *How do aspects of Aesthetic Labour in the cosmetics industry influence the perceptions and expectations of prospective employees?*

Secondly, by combining the two theoretical frameworks, it can be assumed that Aesthetic Labour is valued to a certain degree by applicants when deciding to apply to a company. Reaching the ideal self plays a big role in this assumption, leading to research question number 2:

> *To what extent do job seekers prioritize aesthetic aspects when evaluating potential employers in the cosmetics industry?*

Lastly, it can be assumed that Aesthetic Labour acts as a factor in shaping expectations about work inside the company by portraying a certain lifestyle to the outside. Through past research, it can be assumed that these expectations are not fully met inside the company. This leads to the third research question:

> *How does the perception of a job in the cosmetics industry that is shaped by aesthetic aspects differ from the experience within the company once employed?*

Most existing research on Aesthetic Labour focuses on its direct impact on customer interactions, with little attention given to how it influences perceptions of the corporate brand among potential employees. Furthermore, it is notable that the concept has rarely been connected to the beauty and cosmetics industry, although branding strategies in the cosmetics industry often focus on beauty and aesthetics, glamourizing the use of beauty products through marketing campaigns. This is why it is plausible that the glamorized image of models or beauty retail workers has the same effect on the employer image of the cosmetics industry as the use of Aesthetic Labourers in retail and hospitality. The actual connection of the glamourized view of cosmetics companies shall be researched in depth through the following study.

By combining both literature on EB as well as AL, it can be assumed that Aesthetic Labour may shape the employer brand and thereby impact job seeker perceptions in one of two ways. The employer image may become overly glamourized, attracting many talents that align their own values and wishes with those that are portrayed by the company. Conversely, potential candidates might feel excluded and deselect themselves from a job if they perceive that their appearance does not align with the brand's aesthetic standards. This remains to be researched.

By identifying the connection between Aesthetic Labour and employer branding, this research will contribute to a deeper understanding of how aesthetics influence employment dynamics in the cosmetics industry. It is expected that the findings will feed into both academic discussions as well as practical HR and brand management strategies, particularly in industries where aesthetic factors play a significant role.

4 Methodology

This chapter outlines the methodology employed in this study to explore the connections between Aesthetic Labour and employer branding in the cosmetics industry. To find the connections between all the above-described concepts and answer previously defined research questions, a qualitative approach was chosen, specifically using semi-structured interviews. The chapter is split up into the research design and process, including the sampling strategy and method, the data collection method, and the research process, followed by the data analysis method, and ends with critical factors and limitations in the process of the research.

After having viewed the literature on Aesthetic Labour, it is clear that only little research has been done on office workers, as well as in the beauty and cosmetics industry. Instead, past research has focused mainly on frontline service employees. Connecting Aesthetic Labour to employer branding in

Table 1: Overview Interview Respondents

Interview No.	Age	Gender	Experience at Company (in months)	Division	Active in Company (Y/N)
1	24	F	22	Decorative	Y
2	26	F	24	Skincare	Y
3	24	F	6	Haircare	N
4	24	M	6	Skincare	N
5	24	F	10	Decorative	N
6	25	F	6	Haircare	N
7	26	M	6	Decorative	Y
8	24	F	6	Perfumery	N
9	25	F	6	Decorative	Y
10	25	M	6	Skincare	N
11	22	F	6	Decorative	N
12	23	F	4	Perfumery	Y

the beauty and cosmetics industry can be done by conducting interviews with (former) employees who can compare the external image and expectations to the reality inside the company, especially in office jobs.

4.1 Research Design and Process

To help reach this research objective, a qualitative approach in the form of semi-structured open interviews will be taken. The reason for this approach is due to the topic's very subjective and emotional nature. Concepts like aesthetics can almost not be quantified, nor can general attractiveness, employer attractiveness, or image. All concepts are very emotionally charged and individual to each person, handling complex phenomena like perceptions and personal narratives. This is why a quantifiable suggestive approach would not achieve the desired results within this study. By leading semi-structured interviews, a general guideline will allow the researchers to find connections between the two concepts of employer branding and Aesthetic Labour whilst also giving the interviewee the space for a personal narrative about their experiences. This may also help in gaining insights into experiences that have not previously been considered.

By conducting these interviews, answers to the previously defined research questions will be given.

4.1.1 Sampling Strategy and Method

As the focus of this study is on Gen Z and its seemingly changing values in the labor market, the sample will be chosen according to the following criteria. Interview respondents should be of age but still a part of Gen Z. This means that respondents should be born no sooner than 1995 and be no older than 29 years of age at the time of the interview, as these were all conducted in the first half of 2024. Furthermore, respondents should have relevant work experience in the beauty and cosmetics industry to be able to speak on their accumulated experiences. Lastly, respondents should have no more than three years of professional work experience in total to be able to best remember a time before employment and reflect on their expectations of work in the industry. To conclude, a purposive sampling method was used to select participants who met the criteria. As most (past) employees can speak on their experience in this sector, the search for suitable interviewees should be relatively easy and can be done through the interviewer's network, as well as suitable social media platforms like LinkedIn. Furthermore, by implementing the snowballing method, contacts will be asked for a recommendation of further potential participants who will later be contacted. This way, it can be ensured that there is a good mix of personalities and companies represented to be able to draw a conclusion.

Through this sampling method, a total of twelve fitting interview participants were identified, ensuring a comprehensive set of data for in-depth analysis. Table 1 showcases general information on the interview candidates. All interviewees had a maximum of three years of working experience when entering the company. The respondents all belonged to Gen Z, as they were no older than 29 years of age when interviewed for this thesis. All questioned participants had a past of working for well-known brands from the beauty and cosmetics sector, including decorative cosmetics, haircare, and skincare brands, and were able to give insights into five of the biggest companies in the industry.

Table 2: Interview Guideline

	Values		Expectations	Experience	
General information	**First brand touchpoints**	**Decision to apply**	**Expectations of job**	**Realities in job**	**Aesthetic standards**
Age	First memory of the brand	Influence of company's focus on beauty and aesthetics	Expected work culture	Confirmation of expectations	Personal fulfillment of aesthetic standards
Duration of employment (in months)	Emotions in connection with the brand	Importance of specific aesthetic elements	Expected type of person	Differences to expectations	Origin of aesthetic standards
Category of employment in cosmetics sector		Influence of portrayed lifestyle through company		Differences to employer promises	
		Elements that influenced brand image			
		Comparison to other companies			

As can be seen, especially the gender distribution in this study reflects gender distribution in the real world quite well. Most participants were active in the decorative cosmetics sector, meaning mass consumer make-up brands that are sold at drug stores and perfumeries. Most candidates had a short-term work experience of roughly six months at the company through a timely limited internship, leaving the company afterward. Of those employees who have already left the company, only one would not consider working there again.

4.1.2 Data Collection Method

Data will be collected through semi-structured interviews, allowing participants to share their narratives and experiences. To structure the interviews, an interview guideline was developed, covering three main groups along the research questions to ensure that each question can be answered and no relevant information is missing. The main groups included questions on values, expectations, and experience and were created using a deductive approach, according to Mayring, 2010, using the information that was taken from literature and from those forming assumptions that later led to the research questions. To ensure consistency throughout the interviews, all categories shall be included in every interview. However, depending on the answers of the participants, further questions may be added, modified, or left out. Questions shall allow participants to reflect on their thoughts and emotions towards a certain brand before having been employed there and compare those to thoughts and emotions that re-shaped as soon as they were employed in said company. The following table (see Table 2) shows how questions are categorized and sub-categorized. Each column shows one category of questions that will be asked, and below them, sub-categories in accordance with questions that will be asked.

The full interview guideline has been attached in Appendix A. The interview will begin with questions about initial touchpoints with the brand of the company the interviewee is currently employed in, already leaning into the company's focus on aesthetics and beauty. After that, questions about the interviewee's time of application will be posed to find out how the company's aesthetic focus influenced the decision to apply and how important aesthetic components were in the decision-making process when deciding to take a job at this company over another one, concluding the "values" section of the interview. In the next set of questions, respondents were asked to narrate their expectations after receiving a job offer from the company. This block includes questions about their expectations of the job itself, company culture, well as the type of person that works there. In the last category, the interviewee's current view of the company is compared to their initial impressions to gain an understanding of whether the glamorized view of the company, if there ever was one, to begin with, has changed. This structured approach ensures that all relevant dimensions are explored while providing the flexibility to delve deeper into the interviewees' narratives as opportunities arise.

By categorizing the questions into these three groups and additionally encouraging interviewees to share more than just responses to direct inquiries, the interviews will capture a broad spectrum of experiences, from initial impressions to evolved perceptions. This method will not only confirm the structured research approach but also enrich the data with spontaneous insights, providing a more holistic view of the influence of Aesthetic Labour on employer branding.

4.1.3 Research Process

Of the twelve interviews, eleven were conducted online via Microsoft Teams or Google Meets for flexibility. One of the

interviews took place on the campus of the University of Economics and Business Vienna in a private room, ensuring privacy and a safe place for the former employee to speak about their experiences. Before interviewing the respondents, a few minutes of small talk were held to make them more comfortable. Interviews were recorded with consent, transcribed using automatic transcription software, and proofread for accuracy. Generally, measures were taken to protect the privacy of participants, including the anonymization of all respondents' private data and securing digital files. All interviews were transcribed using an intelligent verbatim transcription method, meaning that filler words like "um" or "uh" were removed to improve readability unless necessary for context. For full data protection, all company names were replaced by placeholders. Initially, all interviews were meant to be held in English. In two of the interviews, however, the language had to be switched to German after the first few minutes, as it was the interview candidates' mother tongue, and they were able to express their experiences much more clearly. For this case scenario, interview questions were carefully translated beforehand to ensure homogeneity. Later, the interview transcripts were translated back to English with careful precision to be comparable to the other interviews.

The interviews had an average duration of 30 minutes, which was enough time for respondents to give insights about their experiences. The shortest interview had a duration of only 20 minutes, while the longest one took 38 minutes.

4.2 Data Analysis Method

To analyze the interviews and gain insights into Aesthetic Labour and employer perception that have not been previously discussed in literature, a deductive approach along the previously defined categories will be most suitable. According to Mayring, 2010, the deductive approach is best used when sorting statements into a before-defined category, allowing for a structured analysis of the interviews. As interviews were conducted rather openly, allowing room for additional questions, the deductive analysis approach can be combined with an inductive approach, if needed, resulting in a mixed-methods qualitative content analysis. To conduct this, the first interviews were thoroughly read, and statements were categorized into three pre-defined categories (values, expectations, and experiences). Any statements that stand out and do not fit into these primarily defined categories will be highlighted to remain open to creating new categories or reiterating the existing ones. Data will be coded and categorized using the MAXQDA software.

4.3 Critical Factors and Limitations of Data Analysis

The concept of Aesthetic Labour concentrates on frontline service workers that represent a brand, thereby shaping its image. The concept is widely applied to the retail and hospitality industries. Known cosmetics brands do work with frontline employees as well, especially at beauty counters in stores like *Douglas* or *Sephora*. It is worth noting, however, that most known cosmetics brands, including *L'Oréal* and its sub-brands or *Henkel*'s brands like *Syoss* and *Schwarzkopf,* are sold in stores like *BIPA* and *dm* in Austria, where the looks of a service employee cannot be controlled by the brands themselves. Besides Aesthetic Labour, marketing campaigns including commercial models of a certain beauty standard, as well as product reviews by trusted role models, may help shape the brand image.

One limitation of this study could include an interviewee's distorted recollection of their experience when working for the brand or when remembering a time before employment at that brand. This is also known as a type of cognitive bias, which occurs when someone is trying to confirm preexisting beliefs by searching for, interpreting, and remembering information that aligns with their beliefs, ignoring contradicting evidence. A cognitive bias is a distorted memory of the past that is not objective and not an accurate recollection of the past. This often occurs when trying to make a decision based on experience – experience is distorted to simplify the decision-making process (Soprano et al., 2024). As recalling emotionally charged memories, like ones connected to role models or favorite brands, cognitive bias may impact the results of this study. Furthermore, employees may not want to speak out on their (current or past) employer in a critical manner. To mitigate this risk, anonymity in all aspects shall be guaranteed to ensure honesty and for interview respondents to feel safe.

In general, the concept of Aesthetic Labour leaves much room for critique, with some authors stating that it "teeters on illegal job discrimination, given the close coupling between appearance and protected categories of race, sex, ability, age, and religion" Mears, 2014, p. 1330. Over the years, a perfect image of companies in the cosmetics and beauty sector has been curated, allowing one to assume that the company stands for and represents only one certain type of person, along with a certain standard of beauty.

The concept of Aesthetic Labour has most definitely gone through several changes over the past years. Through the use of social media, voices of discrimination have gotten louder over the years, "cancel culture" has become big online, and companies are scrutinized for employing only a certain type of person rather than giving equal chances to anyone in the industry. A specific example of this in the retail industry was *Abercrombie & Fitch*, who even went through lawsuits for letting go of people who did not adhere to the company's "look policy," excluding plus-sized shoppers from their clothing lines, as well as claims of looks-based discrimination in their employees (Mears, 2014).

Although it seems as though many of these issues have been resolved or left in the past, a certain image may have still stuck to industries like beauty and retail that are automatically connotated with a certain look. To dive deeper into this critical evaluation, the findings of the interviews will be looked into and discussed over the next chapters.

5 Findings

When conducting the interviews, it became clear that most (former) employees had a similar experience at the company, as well as a certain image that was, at least partially, confirmed. First, values shall be presented, second, expectations, and lastly, experiences.

5.1 Employee Values

When asking respondents about initial brand awareness and perception, most had a very positive image of the brand before applying to work there. This was influenced by market presence, connection to ads as seen on the media, like on television, but also through the connection of an emotional memory or a person they trust being a consumer of that brand. Interestingly, female respondents had a more pronounced first memory of their initial brand awareness, inducing a positive overall response and connotation to the brand. Having role models using and trusting the products of a certain brand builds a first point of contact that curates a positive brand image from an early age. Male respondents more often consciously remember their first touchpoint with a brand to be much later in life, for example, when informing themselves on good, reputable employers or when becoming aware of the products for personal use and interest that developed later in life. However, through the brands' general omnichannel presence, especially on social media, television, and directly at the point of sales (in drugstores like Bipa, dm, Rossmann or perfumeries like Douglas or Sephora), a high brand awareness was given for all respondents. Additionally, it is worth mentioning that almost all interview respondents work or worked for well-known companies or brands. This leads to an overall very positive image of the industry, as well as the employers themselves. Words used to describe the image of their employer include *luxurious*, *aesthetic*, *prestigious*, *sophisticated,* and *professional*. Several respondents mentioned that the products these companies sell symbolize a luxury for them, a treat that they would not purchase or be able to afford every day. Furthermore, respondents often connected the industry to picture-perfect marketing campaigns, through which their image is shaped to be glamorous, shiny, or even glittery. One respondent summarized the image as *"It shines, it glitters, I'd say. A shiny image. I can't describe it better"* (Interview 4, ll.58 f). A summary of all adjectives used to describe the image of the industry can be seen in Table 3, sorted in alphabetical order.

Table 3: Company Image

Interview No.	Company image
1	Aesthetic, exciting, luxurious, picture perfect, prestigious, sophisticated
2	Expensive, high-quality, luxurious, professional
3	Aesthetic, glamorous, luxurious, professional
4	Established, glittery, shiny image
5	Prestigious
6	Clean, positive, shiny
7	Aesthetic, expensive, prestigious, inclusive
8	Authentic, clean, exclusive, modern, high-end, luxurious, progressive
9	Aesthetic, clean, colorful, expensive, luxurious, modern, playful, trendy
10	Innovative, prestigious, reputable
11	Expensive, glamorous, high-glam, luxurious, progressive, shiny
12	Aesthetic, expensive, glamorous, luxurious, modern

The positive image that is intentionally curated by the companies, connected with trust that was induced by role model's use of the products, played a big role when deciding to apply for a job in the industry, specifically at certain companies selling products they already used. Through personal interest and trust that was built towards the brand as a consumer, respondents had a greater urge to "be a part of it," one candidate responding with "*the better it looks, the more glamorous and exciting, the more I personally [. . .] wanted to work there and be a part of that somehow, a part of the company and maybe also the lifestyle they market*" (Interview 1, ll. 200 ff). This statement shows that the image that is portrayed by a company to the outside, including glamour and polished, perfectly curated marketing campaigns, does have an impact on the company's attractiveness as an employer. Through having used the products even before employment, respondents wanted to embody an ideal version of themselves, turning them into a consumer of the products first and later into an employee. Almost all respondents underlined the need for a basic interest in the products and the industry, and this can be watched across all observed companies. The brands' general marketing strategies also made a big impact on their decision to apply, especially when they portrayed a specific image or lifestyle the applicants wished

to achieve. Respondent 8 mentioned, *"I think it was almost the prerequisite that I wanted to apply there because I saw that they somehow radiate exactly what I wanted to be. They have the same image of beauty of aesthetics. My wish was to create a campaign like that and somehow participate creatively and also represent the company in this way"* (Interview 8, ll. 102 f).

Another aspect that was named by several of the respondents was *prestige*. A company's good reputation positively impacts the employer's image, playing a big role in an applicant's decision to apply. This was especially relevant for the male respondents, who initially did not have an emotional connection to the company from being a consumer of the products themselves. However, several females mentioned this aspect additionally to highlight the company's good overall reputation. As stated before, most respondents worked for well-known companies in the industry, which would look good on the CV.

Other aspects that influenced the decision to apply included the portrayal of the company as modern, inclusive, and progressive, an image that has been built up by several companies in the industry over the past few years.

When it came down to deciding between two companies in the cosmetics industry, general efforts of employer branding and attractiveness became relevant in the decision-making process. Here, functional and economic benefits like *professional development*, *pay*, or *geographic proximity* became crucial for the decision. Mainly, however, several respondents did talk about colleagues (interns) whose main motivation to apply to the job was the company's *prestige*, one respondent stating: "*People just want to go there because it says something about themselves*" (Interview 5, l. 307). Several respondents did mention turning down jobs in the food industry or for other FMCG companies, implying that they valued the image and products inside the beauty and cosmetics industry more greatly.

This concludes the findings of the first section of the interview that focused on the values of (former) employees. While functional and economic benefits do remain important for the application decision, it is clear that very many applicants are pulled in by the company's positive image, prestige, and portrayal of a lifestyle to the outside.

5.2 Expectations

In the next step, the respondents were asked about their expectations of work inside the company, specifically how they imagined company culture and the type of person that worked there. Respondents were not explicitly asked to describe the way employees would look, but rather, which type of person they expected to work in the industry. The question was purposefully posed openly to see whether looks played an important role in the expectations of a future employee. Most respondents did describe the way they imagined a potential colleague to look like. One respondent even went as far as describing a persona. *"It would be a female person. Middle-aged. Not too old, but also not super, super young. [. . .] I think young people bring in this new spirit, which always represents this novelty. [. . .] Very modern and innovative in terms of mindset. And also, in terms of appearance, the way the person looks physically. Probably more sporty. Also, this clean girl aesthetic. Simple, clean. 'I take care of myself'. 'I have the need to take care of myself'"* (Interview 6, ll. 178 ff). There was a very clear image of a person working in this industry. Most other respondents also had a clear image in their minds when describing what they expected their future colleagues to look like. Respondents mentioned that they thought employees would be mainly *female*, have a *beauty affinity*, be *chic*, *fashion- and cosmetics-conscious*, *beautiful*, *"styled to a T,"* fit into a *European standard image of beauty*, described as blonde hair and white skin. They would always be *put together*, very *neat*, and have *perfect skin*. This description oftentimes fits the persona the companies use to advertise their products and brands. "*You think that they really embody what the brand embodies*" (see interview 2, ll. 190 f). It is noticeable that all respondents who mentioned the looks of employees used very positively connotated words to describe these looks. In Table 4, adjectives that were used to describe expected employees are cumulated.

However positive the expected looks of prospective employees were, the expected personalities of employees differed and included positive as well as negative characteristics that employees might have. Expected positive characteristics included words like *hard-working*, *ambitious*, *innovative*, *professional,* and *creative*. However, in this category, many respondents included negatively connotated words like *snobby*, *strict*, *arrogant*, *superficial*, *mean*, *elbow-mentality*. One respondent even explicitly mentioned being afraid of these traits, stating, "*I was scared that the people working there were also kind of like, how do you say it, a bit snobby*" (Interview 2, ll. 125 f).

A few respondents also mentioned having a prejudice in thinking the type of person who purchases the products would also work for the brand, which is where adjectives like *wealthy*, *elitist*, *academic,* or *snobby* came into play, being underlined by statements like "*I assume that the people who buy the brand also work for the brand*" (Interview 2, ll. 180 f).

Some respondents spoke about the origin of this image they had of an employee in the beauty and cosmetics industry. Here, it is evident that pop culture, movies, and television shows, as well as the company's own branding and marketing campaigns, shape general expectations. Some respondents named movies like *The Devil Wears Prada, Ugly Betty,* or *Mean Girls*, where employees that work in this industry are portrayed the way they described them: having a put-together look on the outside but having rude and condescending character traits. Others mentioned their role models having a decisive factor in shaping their image of a person inside that industry, labeling the women seen on popular television shows like *Germany's Next Topmodel* as *ungrateful* and even *bitchy*; an image that has spilled over to the respondents.

Respondents were also asked about their expectations of the general working culture and the job itself. It was no-

Table 4: Expectations

Interview	Expected Characteristics	Expected looks
1	Hard-working, performance-oriented, ambitious, motivated	–
2	Snobby	Luxurious, wealthy, upscale
3	Multi-cultural, innovative, professional	Picture-perfect, beauty affinity, eye on aesthetics, female-dominated
4	Elitist, academic	Sparkling, formal, chic, fashion-& cosmetic-conscious, female, young
5	Communicative	Embody aesthetics and beauty, girly, chic
6	Modern & innovative mindset	Clean, shiny, well-groomed, beautiful, female
7	Diverse, strict, formal, superficial, arrogant	Aesthetic, put together
8	Like-minded, creative, driven	Standard European beauty image, uniform
9	Professional, creative, outgoing	Colorful, neat, wearing makeup, outgoing, eye for aesthetics
10	Fancy, extroverted	"Marketing girlies", fancy, put together
11	Mean, elbow-mentality	Ideal of beauty, perfect skin, put together, dressed up, chic, picture-perfect, blonde, white
12	Hard-working, ambitious, motivated, arrogant	Beautiful, chic, put-together

ticeable that the positive image they had of the company beforehand also translated to the expectations about the job. Respondents were very excited to work at a well-known company, expecting them to be modern employers with flat hierarchies, well thought-out, professional, serious, and dynamic processes. Respondents had hoped to be a part of the strategic creation of campaigns and be able to creatively give individual input. This shows how positively the company image that was curated before even applying, through marketing campaigns and word-of-mouth, as well as external representation through employees, also impacts the expectations in the work itself.

To conclude the section on expectations, it can be said that expectations were set by the image that already existed inside the prospective employees' minds, mainly shaped by media.

5.3 Experience

To be able to compare these initial expectations to the reality of working inside the company, interviewees were then asked to compare their expectations to their experience. Here, results can be split up three-fold. There were positive and negative changes in perception, as well as confirmations in expectations. In general, it was interesting to see that respondents who, at the time of the interview, were still currently employed at the company mentioned far more positive aspects rather than negative ones.

Positive changes were perceived, especially concerning the characteristics of colleagues and other employees. Expected negatively connotated characteristics like *arrogance* or *snobbiness* were retracted and replaced by words like *human*, *approachable*, *diverse*, *creative*, *unique*, or *warm*. Colleagues inside the team were enthusiastic about the products, through which a positive work environment was created. Colleagues connected through personal interests and joy in beauty, were proud to work in the industry, and showed themselves to be very supportive towards others. Many respondents did expect the employers to be rather modern and not conservative. However, even those who did not think about this before mentioned the companies being very modern, progressive, and colorful. One respondent said, "*We are moving away from the [. . .] socially constructed image of beauty, and there is simply a lot of progress*" (Interview 1, ll. 87 ff) concerning positive developments in the direction of consumer needs, especially in terms of sustainability, diversity, or in aiding disabled people in recent years. Through an overall positive experience, many of the respondents still feel an emotional connection to the brand even after employment, opting for the products of the brands they worked for or feeling a sense of pride when seeing the products in stores.

Besides a very positive and uplifting work environment in general, some aspects were perceived more negatively than originally expected. Many of the respondents were disappointed in the strategy, seeing many things pre-dominated

through the parent company or simply repeating old ideas, taking away strategic and creative elements from the job. Because of this, one candidate even said: "*I kind of already lost hope a bit that it could really be my dream job*" (Interview 8, ll. 191 f). Some respondents who completed an internship mentioned a relatively flat learning curve, repetitive tasks, and little innovation inside the team. In general, processes were described to be slow rather than dynamic, leading to overtime. On the other hand, when there was innovation, there was very much pressure to be innovative, leading to employees crying or having anger outbursts. Market research results were partially even faked to secure internal budgets, and processes were not as thought out as expected. Through this, one candidate stated that "*this glamorous and professional image that I had of many brands has disappeared a bit*" (Interview 3, l. 173). Another aspect that was mentioned here was how competitive the work environment in general was. Although inside the teams, a positive overall attitude ruled, there was constant pressure and competition to rise in hierarchy or to secure a full-time job after an internship. There is only little negotiation leverage, as the companies know very well that there are hundreds of applicants who will work for them under their own conditions. Through this, especially in larger companies, work could seem impersonal, and other colleagues could even seem a bit arrogant.

A confirmation of expectations was seen, especially concerning the look of employees who work in the beauty and cosmetics industry. As expected, colleagues and other employees in the industry had a general interest in beauty and luxury, expressing their interest through clothing and makeup, achieving a very *upscale*, *chic*, *perfectly coordinated* look. As one respondent stated, "*You have a certain expectation which people go to [company] and are interested in it. And that's how it is*" (Interview 4, ll. 184 f), mentioning that you expect employees to be fashion- and cosmetics-conscious. Another respondent also confirmed this by stating employees personify the "*core concept of aesthetics and beauty. [...] That's what the people on average embodied very strongly*" (Interview 5, ll. 169 f). Throughout the interviews, it was noticeable that many respondents mentioned a very uniform, homogenous look. Although some respondents mentioned a development towards more diversity, many stated that everyone has a certain look, and many follow a classic (European) beauty ideal – as one respondent stated: "*It definitely is one certain type of person that works there, we're very similar, we all look the part*" (Interview 12, l. 203 f). This was so noticeable throughout the different companies that many even asked themselves whether this was purposefully or intentionally done to maintain a certain image to the outside. Beauty standards were noticed especially in customer-facing roles, like in sales or marketing and event departments – departments in which employees play a representative role to the outside. One respondent connected this realization to the origin of their image of the industry beforehand, stating: "*They always take as representatives people who best address the potential employee in their opinion. And so you can see again, okay, those are the type of people who work there. And that spreads, that goes through at some point. Once you've seen it once, you probably transfer it directly to several people*" (Interview 6, ll. 199 ff). Several respondents can imagine that this is a selection bias that already takes place during the employment process, during interviews, or even through pre-defined criteria that AI goes through when evaluating CVs and motivational letters in the company's web interface. Whether or not companies evaluate applicants by their looks, which would indicate the targeted use of Aesthetic Labour, could not be confirmed. However, many respondents indicate in their statements that there is a preference, one respondent assuming that "m*aybe in marketing you have a lot of touchpoints to customers, especially to end-consumers. Maybe there's a reason why and they do it on purpose*" (Interview 11, ll. 308 ff). All statements with assumptions about a selection bias can be found in Appendix B.

Fulfilling a certain aesthetic standard is not only a part of the selection process but can additionally be observed inside these companies. Although maintaining a certain look is not obligated by the company, a certain internal pressure is noticed by the respondents, stating they want to look nice, wear makeup, and dress chic to fit in. Pressure is noticed by most respondents, especially females, to adhere to a beauty standard, put additional thought into outfit choice, and to represent the brand, even act as a brand ambassador for the company. Male respondents did not notice this pressure as much, comparing it to a level of professionalism that would be found in every industry. According to many of the female respondents, the entire company culture is built up around acquiring and maintaining a beauty standard, and even though individuality is celebrated more and more, there "*is a certain culture here, and you have to fit in. And if you don't fit in, then it's not for you to work there*" (Interview 5, ll. 180 f). Some respondents even mentioned that it was noticeable the higher in the hierarchy employees were, the more they fit into a classical standardized beauty image, indicating that, to be promoted, you do have to fit into the company's standardized look. Some respondents did discuss the changing image of beauty inside the industry, stating that it was oftentimes the older generations that still hold on to an outdated image of work inside the industry, their mindset still being different and holding on to the etiquette that was established when they began working. "*For them, it was just a lot more classic, and if you dressed less chic, you might have been perceived as less*" (Interview 9, ll. 305 f).

One respondent touched more deeply on this changing image of the beauty industry in general, stating that it is a process. They are excited to "*see how this image will change in general as soon as the younger generation, even younger than me, enter the job market and carry their own image of the beauty industry with it. I think they are already growing up with a completely different image*" (Interview 11, ll. 237 ff). Whether or not this will change intentional or unintentional selection biases remains to be discovered in the future.

To summarize the findings about the experience inside the company, it can be said that some respondents were

blinded by the industry's shining image and disappointed when seeing behind the scenes. The perfect image that is portrayed to the outside not only shapes product brands but also the employer brand. By perceiving a very clean and perfect image, applicants may automatically assume that the processes behind the final product also run more smoothly, leading to disappointment in the long run. It is clear that aesthetic standards are very much a part of the work culture inside these companies, leading to internalized pressure to fit in. To conclude, one respondent mentioned that the way a company portrays itself to the outside world, both through employer branding as well as through general product marketing, will naturally attract applicants who identify with those products and see themselves in that lifestyle (see interview 1). Another respondent's reply adds to this statement: in the experience of respondent 2, there is a very cohesive, uniform standard concerning looks in departments that work a lot with consumers to keep up a certain image. To credibly sell a brand to the consumer, they should "fit in" to the brand and radiate the brand's image and values, which, in a superficial industry, includes fitting the look. "*When you work for the brand, use the brand, AND embody everything they promise, so you don't have pigmentation spots, you don't have wrinkles, then, of course, the whole picture becomes more credible. And of course, that influences the brand image, and the trust people have in the products enormously*" (Interview 2, ll. 314 ff), indicating a reason why beauty and cosmetics companies may intentionally be applying Aesthetic Labour practices.

6 Discussion

In the beginning of the thesis, three research questions were posed, whose aim it was to find the connection between Aesthetic Labour and employer branding. The first research question aimed to find out, *how Aesthetic Labour practices influence the perceptions and expectations of (prospective) employees*. Through the interviews, it became clear that, for all respondents, brand image was not only formed by classic EB strategies. For many respondents, aspects like brand knowledge and loyalty, as well as identification with the products influenced the image of the company and brands more than standard EB practices – especially for females, that were exposed to the brands and products from a very young age. By seeing consumers in perfumeries or watching role models like their mothers use and trust the products, a positive image and connection was automatically formed from a young age. This fits to the research of Pounders et al., 2014, who also mentioned rising trust in a brand, the more aligned it is with customers interests. Furthermore, models in beauty campaigns portrayed an ideal that respondents mentioned they wanted to follow and be a part of (Guthrie et al., 2008; Kim & Sullivan, 2019). This shows that many of the female respondents were consumers of the products first, before wanting a look behind the scenes and deciding to apply. Male respondents often times had less of an emotional connection to the companies and brands, looking for things like prestige or a "big name" for the CV. For them, brand loyalty grew more strongly during and after employment (Cutcher & Achtel, 2017). It is true, however, that the image that is created through product marketing campaigns influences the expectations of the respondents on the type of person who works in a cosmetics firm (Barnett et al., 2006; Collins & Stevens, 2002). When asked to describe the type of person who could work in a cosmetics company, most named traits that adhere to the classical beauty standard. Many respondents had a very clean, professional, perfect image of beauty and cosmetics companies. This could be due to the clean, high-gloss campaigns with models that show no or only little signs of imperfection. For years, the cosmetics industry used this classic approach to sell a product promise. Meanwhile, more and more campaigns include unedited models, with pigmentation spots or wrinkles, utilizing the *actual self* to speak to consumers (Malär et al., 2011). Still, it seems as though the old image prevails in the minds of many. It is possible, that the image that is formed when the first touchpoint with the brand is created, often at a young age, through watching role models, remains engrained into people's minds and is hard to change. Even with cosmetics companies' recent efforts to show more inclusion and diversity in marketing campaigns, a very high-glam, picture-perfect image remains.

To conclude, Aesthetic Labour practices, including the hiring of employees due to their looks or utilizing a certain type of model in marketing campaigns do have an impact on brand and therefore employer image. As Luoh and Tsaur described in their research paper published in 2023 on the hospitality industry, a parallel can be drawn to the beauty industry, where the assumption can be confirmed, that Aesthetic Labour practices influence both brand and employer image.

The second research question aimed to find out, *to what extent job seekers prioritize aesthetic factors, such as brand image and lifestyle portrayal, when evaluating potential employers.* As it was previously presumed, it is safe to say that the portrayal of an aesthetic look and lifestyle through employees as well as marketing campaigns have an effect on a company as an employer and the image of the workplace (Luoh & Tsaur, 2023). Here, many respondents, as expected, mentioned wanting to be a part of what the brands portrayed and wanting to look behind the scenes of the picture-perfect image they had from the outside. Mainly female respondents were pulled by the image and lifestyle, and even more so if they already used and trusted the products. Often times, young girls get acquainted with brands through people they trust. This could be their mom, aunt, or other role models. Surely, this too has a connection to the ideal self – by seeing the products that trusted role models use, emotional values like their own interest, trust and loyalty in the brand grow (Lies, 2021). This familial connection played a notable role in shaping respondents' perceptions and trust in the brand before even considering it as an employer. Male respondents, especially when they had no touchpoints to the products at all, did not consider lifestyle portrayal or other aesthetic factors as part of their decision to apply and take the job. Again,

rational factors like functional and economic benefits (see chapter 2.1.1; Ambler and Barrow 1996) were weighed more strongly. Generally, symbolic elements of the employer were weighed more strongly than instrumental elements (Reis et al., 2017) – these became more relevant when deciding between two companies in the cosmetics industry. In this case, symbolic elements can be viewed as the company and employer image that is shaped by Aesthetic Labour practices. It was also interesting to see, how few respondents mentioned diversity and inclusion as a reason for company attractiveness. Apparently, the aesthetic image that was portrayed for so many years still prevails and changing a mindset will take longer than trying to change a company's image. To conclude, Aesthetic Labour practices along with its consequences did play a role when respondents decided to apply.

Lastly, it was researched, *how the perception of a job in the cosmetics industry, that is shaped by aesthetic aspects differs from the experience within the company, once employed*. In the beginning, when exploring respondents' image before *employment*, relatively many negatively connotated adjectives, like snobby, superficial and arrogant were named to describe expected characteristics. However, many of these descriptions were redefined, as soon as the respondents started working inside the company and getting to know colleagues. Respondents mention noticing that a lot of effort is made inside the company to upkeep a certain image to the outside. A few respondents mentioned that even inside the office building, the entry halls are usually very glamorous. However, a few floors up, the office building could belong to any other company. It is as though the entry hall is used as a showcase for the company, to impress potential clients, including flagship stores, luxurious coffee bars or photo walls. It is noticeable that in departments, where employees have much contact with (potential) customers, a certain look and aesthetic is sustained, to keep up a certain image (Mears, 2014). Here, it really becomes evident, that cosmetics companies are not only selling beauty enhancing products, but they are also selling a lifestyle in every single area (Cutcher & Achtel, 2017). Besides keeping up the image, these companies operate just as any other business does internally. When asking respondents about their experience inside the company, the consensus on the *type of person* working there was very positive. They emphasized, how well you can develop your career based on interests, being motivated by very driven colleagues, and adjectives used to describe employees of those companies turned positive, including more words like *ambitious*, *engaged* and *enthusiastic*. This really shows that some of the characteristics that are associated with a certain "look" are much more superficial and negatively connotated than the reality shows.

It is true, that a certain appearance was upheld in all companies. Although there were no official rules or a dress code, most respondents did feel pressured to maintain a certain standard. This is mainly because of an internal pressure – most employees that work for these firms have a very high interest in beauty, so they do showcase it to the outside. This in turn creates an internal pressure for other employees and, although they are not forced to, they still dress a certain way and maintain a certain look to fit in. Especially in leading managerial positions, employees embody the epitome of what the beauty industry stands for, which shows, that if you want to climb the career ladder, you should embody what the beauty industry stands for on a daily basis. As maintaining a certain look needs to be conducted outside of work, this pressure to do so may very much turn negative, into a psychological burden (Entwistle & Wissinger, 2006). To conclude, employees do have a preference for luxury, beauty and aesthetics and showcase that in their own styling, while still having humane, friendly, open-minded mindsets.

Aesthetic Labour is not implemented as strictly in office working cultures, and not regulated by the employers themselves. This shows the main difference to frontline workers, where employees are purposefully picked based on their looks (Luoh & Tsaur, 2023; Mears, 2014; Warhurst et al., 2000). However, several respondents mentioned, that there may be a bias on HR side, as (almost) all employees working for these companies portray a uniform image of beauty to the outside (Vonk, 2020). Would someone, that is not interested in maintaining their looks, be offered the same job opportunities? The strategic implementation of Aesthetic Labour as employer branding strategy cannot be confirmed. It can be assumed, however, that openly communicating the implementation of Aesthetic Labour practices, especially in a world that is values inclusion and diversity more and more, would be more harmful to a company than it would do good (Cutcher & Achtel, 2017).

Concerning the working culture inside these companies, several employees were disappointed to look behind the scenes of a company that, to the outside, is portrayed as picture perfect. Internal processes do not work as good as imagined, there was a lack in development of strategy, there was less structure than expected, tasks became repetitive, and the learning curve (especially within internships) became relatively flat. One respondent stated, that "*this glamorous and professional image that I had of many brands has disappeared a bit*" (Interview 3, ll. 173 f), because they now know how they are managed internally. This re-shapes the employees' employer image, which is shaped by the combination of external perceptions and internal realities (Sivertzen et al., 2013).

To answer research question number three, the perception that employees had of companies in the cosmetics and beauty industry does differ from the reality inside the company concerning work inside the company as well as the characteristics of employees. However, expectations concerning the look of employees can be confirmed, confirming the assumption that employers may strategically or unintentionally be curating a uniform look throughout the company (Vonk, 2020).

In general, based on the analysis of the conducted interviews, it can be said that especially media including television, movies, and social media, marketing campaigns, as well as employees themselves shape the brand image. By portraying this image of beauty and aesthetics to the outside, people

that are interested in these product categories, often times already including the products of these brands in their private lives to become a version of their ideal self (Kim & Sullivan, 2019), are more likely to apply here. The EB practices, that are implemented by companies in the beauty and cosmetics industry seem to work, as all respondents described very similar personas, when speaking of their colleagues.

The image that applicants have of a company is generally influenced by marketing campaigns and personal relations, which goes along with the branding process of Backhaus and Tikoo (2004, see chapter 2.1.1). By winning over consumers first, they will most likely obtain a positive image of the company itself, and want to be a part of it, turning consumers into applicants and later employees. The respondents that were interviewed for this study did all work for very reputable employers, that, through branding strategies, improved their image as an employer. The good employer image was important to all respondents; however, male respondents mentioned it more predominantly. This may be due to their interests in the brands and products themselves being lower. As stated in chapter 2.1.2, factors that positively impact employer image include innovativeness, CSR initiatives and the effective communication of brand values. Especially the latter is communicated very strongly by companies in the beauty and cosmetics industry, to allow strong employee identification with the company and its brands (Brannan et al. 2015).

The companies' general attractiveness was also mentioned in many of the interviews. When comparing Berthon and colleagues' EmpAt model to the statements of the respondents, it becomes clear that interest value, social value, and development value were very interesting to the respondents in the beauty and cosmetic industry. Only two of the respondents mentioned economic value, showing that it is rather unimportant in the industry. One respondent stated that "*as an applicant, [you] have bad cards to ask for much more payment in the application process. Because they know very well there are 250 other people in line who absolutely want this job. [...]. Just because it's an aesthetic company*" (Interview 5, ll. 312 ff). Here, you can see how unimportant economic value is at first when the main goal is to enter the company. Ružić & Benazić also discovered this in their study, stating that material working conditions are less important for employer attractiveness Ružić and Benazić, 2023, see chapter 2.1.3. Application value was something that the respondents did not mention when speaking of their attraction to the company before employment. However, it quickly became clear that, even in an aesthetic company, it is important to have tasks that are not too repetitive. As stated in Chapter 2.1.3, Gen Z especially values a fun work environment, a positive team environment as well as a supportive relationship with colleagues and superiors. The interviews partially confirmed this, with respondents highlighting the positive and familial work environment within teams, making these companies attractive to them (Lassleben & Hofmann, 2023). This does, however, not explain why respondents applied to the companies in the first place, as this was not the image they had of companies within the beauty industry beforehand.

All respondents confirmed that Aesthetic Labour was not an official prerequisite by the employer, unlike in the retail industry. However, they do still have very much control over the type of person they employ, and employees (respondents) do notice a very cohesive look inside the company, which might imply either a bias on the HR side or internal, not communicated aesthetic standards. It does make sense for large companies that sell products that enhance beauty to work with ideal images, especially to win over consumers with low self-esteem (Malär et al., 2011). Through marketing campaigns, brand ambassadors, and even employees that all follow one image, consumers will be more likely to purchase these products, feeding into their own ideal selves (Lies, 2021). The coherent portrayal of brand images throughout the cosmetic and beauty industry creates an image of the self that consumers want to follow. Ultimately, this even feeds into the desire to be employed at the company. Many respondents mentioned this, stating things like "*I really wanted to be a part of it*" (Interview 2, ll. 75) or "*I think it was almost the prerequisite that I wanted to apply there because I saw that they somehow radiate exactly what I wanted to be. They have the same image of beauty, of aesthetics*" (Interview 8, ll. 101). Cutcher and Achtel, 2017 touch on "selling a lifestyle" when hiring employees, an aspect that was mentioned several times and, therefore, underlined through this study. Whether or not this leads to self-deselection of other candidates could not be researched (see Cutcher and Achtel 2017), as all respondents did have a job in the industry.

It can be said that the portrayed and advertised image somewhat distort the reality of work in a company. Automatically, recipients assume that the perfect portrayal of the industry can be transferred to the way of working inside the company, which is often untrue.

7 Practical Implications, Limitations and Directions for Future Research

This study mainly shows that all that glitters is not gold. Employer branding through the use of Aesthetic Labour, aesthetic campaigns, and brand ambassadors is a great way for companies to curate a "picture-perfect" image of themselves to applicants. As it became clear, the image that employers in the industry curated does not always match up with the realities inside the companies. For young applicants, it may be interesting to look past this initially curated image and rank employers based on attractiveness factors that are more important to them.

7.1 Practical Implications

This study is interesting, especially for HR specialists working in the beauty and cosmetics industry and focusing on a more diverse representation inside the company. As it is clear that Aesthetic Labour practices can lead to negative effects, like self-discrimination or deselection from specific

jobs, HR specialists inside the industry could ensure that more inclusive, diverse, and unbiased employer branding practices are implemented within the company. To prevent selection biases when employing new personnel, companies in the beauty industry may consider outsourcing their hiring processes to unbiased third parties to ensure equal opportunities for all applicants, independently from their looks.

Another topic that was briefly touched on was the psychological burden that may arise from the pressure of fitting into the look that is predominant in the industry. Managers inside these companies should, primarily, gain a broad understanding of the impact of aesthetic standards and, going from that, implement employee well-being programs that support things like mental health and job satisfaction.

7.2 Limitations and Directions for Future Research

Limitations of this study include the selection of the sample, which was briefly discussed in the methodology section of this paper. The sample was predominantly female, which reflects the gender distribution in the industry but may limit the understanding of Aesthetic Labour practices from a male perspective. Furthermore, the cultural background of the respondents was very similar, likely influenced by Western beauty standards. This limitation means that the findings might not be entirely applicable to other regions with different cultural norms and aesthetic standards. In future research, a larger sample size, including a more balanced gender representation as well as a bigger geographical distribution, could be taken into account.

Respondents were asked to recall memories and experiences from before their employment, which could lead to a retrospective cognitive bias. This retrospective approach may introduce bias, as memories can be selective and influenced by subsequent experiences. Participants might have unintentionally omitted or altered details over time. A future study could focus on interviewing two focus groups, one that has not been employed in the industry yet, and comparing those findings to a sample group that is currently employed in the industry. Furthermore, the study focused on employees with relatively short-term experience (maximum of three years) and included a high proportion of interns. This may limit the understanding of how Aesthetic Labour practices impact long-term employees and those in higher-level positions within the organization, which could also indicate an idea for future research focusing on the long-term impact of Aesthetic Labour practices.

In general, it was noticeable that interview respondents who were currently still employed in the company they were speaking about utilized more positive language to describe their experience, being more careful to talk about confidential details. This also represents a limitation in this study that may have been prevented through additional measures to make respondents feel safe and comfortable in the interview setting.

Additional directions for future research include conducting expert research with HR professionals to uncover whether there is a strategy behind the assumed selection bias that was found in this study. Furthermore, to be able to effectively implement employee well-being strategies, the effects of Aesthetic Labour on employee (un-) well-being, including aspects like job satisfaction or aesthetic pressure, could be conducted. This way, employee well-being programs can be implemented more precisely.

By addressing the practical implications and pursuing further research in the area, companies in the beauty and cosmetics industry can ultimately enhance their employer branding strategies, create a more inclusive work environment, and improve overall employee satisfaction and performance.

8 Conclusion

All in all, it can be said that Aesthetic Labour does play a significant role in the office work environment of cosmetics companies. Investigating the influence of the concept on employer branding within the cosmetics company has yielded significant insights into how indirect factors, beyond traditional employer branding strategies, shape the perceptions and expectations of prospective employees. This study highlights that brand image and loyalty, often established through early exposure to products through role models or family, as well as marketing campaigns, play a crucial role in forming an attractive employer image. The research underlines the role of Aesthetic Labour in the cosmetics industry, revealing that it not only affects customer-facing roles but is also applicable to internal organizational culture. The portrayal of an ideal aesthetic and lifestyle through marketing and employee appearances significantly influences job seekers' decisions, particularly among those already familiar with the brand. However, this study also reveals a discrepancy between the external, glamourized image and the internal realities of working within these companies, with some employees experiencing a mismatch between expectations and actual working conditions. Contrary to companies in the fashion and retail industry, cosmetics companies do not openly recruit only "good-looking" personnel, leaving them out of controversial discussions in the media.

In Generation Z, there is a noticeable rise in the demand for corporate social responsibility, particularly in terms of diversity and representation. Many respondents observe an increase in diversity-focused and inclusive marketing campaigns from their employers. Although this shift is viewed positively, many employees still report a sense of uniformity regarding the employees within the company. This indicates that a changed external image does not necessarily reflect internal change or that internal change takes far longer to implement than external marketing campaigns. Furthermore, this finding suggests that while Aesthetic Labour may be exclusionary, it does not significantly harm the employee's experience within the company or, more importantly, a company's reputation as long as a different, more inclusive image is portrayed outwardly.

In conclusion, this thesis contributes to a deeper understanding of the complex dynamics between Aesthetic Labour and employer branding in the beauty and cosmetics industry. The findings of this study hold valuable implications for both academic discussions and practical HR and brand management strategies, particularly in industries where aesthetics play a significant role. All in all, Aesthetic Labour represents a darker side of employer branding that has not been researched in this context before.

References

Alizadeh, L., Liscio, M. C., & Sospiro, P. (2024). The Phenomenon of Greenwashing in the Fashion Industry: A Conceptual Framework. *Sustainable Chemistry and Pharmacy*, *37*, 101416. https://doi.org/10.1016/j.scp.2023.101416

Alnıaçık, E., & Alnıaçık, Ü. (2012). Identifying Dimensions of Attractiveness in Employer Branding: Effects of Age, Gender, and Current Employment Status. *Procedia - Social and Behavioral Sciences*, *58*, 1336–1343. https://doi.org/10.1016/j.sbspro.2012.09.1117

Ambler, T., & Barrow, S. (1996). The Employer Brand. *Journal of Brand Management*, *4*(3), 185–206. https://doi.org/10.1057/bm.1996.42

Backhaus, K., & Tikoo, S. (2004). Conceptualizing and Researching Employer Branding. *Career Development International*, *9*(5), 501–517. https://doi.org/10.1108/13620430410550754

Baldessarelli, G., Stigliani, I., & Elsbach, K. D. (2022). The Aesthetic Dimension of Organizing: A Review and Research Agenda. *The Academy of Management Annals*, *16*(1), 217–257. https://doi.org/10.5465/annals.2020.0198

Barnett, M. L., Jermier, J. M., & Lafferty, B. A. (2006). Corporate Reputation: The Definitional Landscape. *Corporate Reputation Review*, *9*(1), 26–38. https://doi.org/10.1057/palgrave.crr.1550012

Berthon, P., Ewing, M. T., & Hah, L. L. (2005). Captivating Company: Dimensions of Attractiveness in Employer Branding. *International Journal Of Advertising*, *24*(2), 151–172. https://doi.org/10.1080/02650487.2005.11072912

Biswas, M. K., & Suar, D. (2016). Antecedents and Consequences of Employer Branding. *Journal of Business Ethics*, *136*(1), 57–72. https://doi.org/10.1007/s10551-014-2502-3

Butler, C. (2014). Wanted – Straight Talkers: Stammering and Aesthetic Labour. *Work, Employment and Society*, *28*(5), 718–734. https://doi.org/10.1177/0950017013501956

Campbell, C., Sands, S., McFerran, B., & Mavrommatis, A. (2023). Diversity Representation in Advertising. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-023-00994-8

Collins, C., & Stevens, C. K. (2002). The Relationship between Early Recruitment-Related Activities and the Application Decisions of New Labor-Market Entrants: A Brand Equity Approach to Recruitment. *Journal of Applied Psychology*, *87*(6), 1121–1133. https://doi.org/10.1037/0021-9010.87.6.1121

Cutcher, L., & Achtel, P. (2017). "Doing the Brand": Aesthetic Labour as Situated, Relational Performance in Fashion Retail. *Work, Employment and Society*, *31*(4), 675–691. https://doi.org/10.1177/0950017016688610

Dassler, A., Khapova, S. N., Lysova, E. I., & Korotov, K. (2022). Employer Attractiveness from an Employee Perspective: A Systematic Literature Review. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.858217

Die Presse. (2024). Doku über Brandy Melville: Wer zunimmt, fliegt raus. *Die Presse*. https://www.diepresse.com/18379534/doku-ueber-brandy-melville-wer-zunimmt-fliegt-raus

Dion, D., & Tachet, B. (2019). Dynamics between Market Categories: A Study of the (In)Visibility of the Plus-Size Fashion Market. *Recherche Et Applications En Marketing*, *35*(1), 62–83. https://doi.org/10.1177/2051570719863648

Edwards, M. J. A. (2010). An Integrative Review of Employer Branding and OB Theory. *Personnel Review*, *39*(1), 5–23. https://doi.org/10.1108/00483481011012809

Entwistle, J., & Wissinger, E. (2006). Keeping Up Appearances: Aesthetic Labour in the Fashion Modelling Industries of London and New York. *Sociological Review*, *54*(4), 774–794. https://doi.org/10.1111/j.1467-954x.2006.00671.x

Fonrouge, G. (2023). Abercrombie & Fitch, Ex-CEO Sued over Sex Abuse and Trafficking Accusations. *CNBC*. https://www.cnbc.com/2023/10/27/abercrombie-sex-abuse-lawsuit.html

Frisby, C. M. (2019). Black and Beautiful: A Content Analysis and Study of Colorism and Strides toward Inclusivity in the Cosmetic Industry. *Advances in Journalism and Communication*, *07*(02), 35–54. https://doi.org/10.4236/ajc.2019.72003

Guthrie, M., Kim, H., & Jung, J. (2008). The Effects of Facial Image and Cosmetic Usage on Perceptions of Brand Personality. *Journal of Fashion Marketing and Management*, *12*(2), 164–181. https://doi.org/10.1108/13612020810874863

Hiller, V., & Verdier, T. (2014). Corporate Culture and Identity Investment in an Industry Equilibrium. *Journal of Economic Behavior & Organization*, *103*, 93–112. https://doi.org/10.1016/j.jebo.2014.04.006

Kim, Y. K., & Sullivan, P. (2019). Emotional Branding Speaks to Consumers' Heart: The Case of Fashion Brands. *Fashion and Textiles*, *6*(1). https://doi.org/10.1186/s40691-018-0164-y

Lassleben, H., & Hofmann, L. (2023). Attracting Gen Z Talents: Do Expectations towards Employers Vary by Gender? *Gender in Management: An International Journal*, *38*(4), 545–560. https://doi.org/10.1108/gm-08-2021-0247

Lies, J. (2021). Aesthetics Rising from Beauty to Reputation Management. *Corporate Reputation Review*, *24*(2), 105–115. https://doi.org/10.1057/s41299-019-00094-w

Lievens, F., & Slaughter, J. E. (2016). Employer Image and Employer Branding: What We Know and What We Need to Know. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*(1), 407–440. https://doi.org/10.1146/annurev-orgpsych-041015-062501

Loftus, G. (2013). Leading by Going After the Cool Kids. *Forbes*. https://www.forbes.com/sites/geoffloftus/2013/05/22/leading-by-going-after-the-cool-kids/?sh=33b269cefc7e

L'Oréal. (2020). Strategie und Modell. https://www.loreal.com/de-de/germany/press-release/group/employer-reputation/

Luoh, H., & Tsaur, S. (2023). Looking Good and Sounding Right: A Scale Development of Aesthetic Labour in the Hospitality Industry. *Journal of Hospitality and Tourism Insights*. https://doi.org/10.1108/jhti-12-2022-0598

Malär, L., Krohmer, H., Hoyer, W. D., & Nyffenegger, B. (2011). Emotional Brand Attachment and Brand Personality: The Relative Importance of the Actual and the Ideal Self. *Journal of Marketing*, *75*(4), 35–52. https://doi.org/10.1509/jmkg.75.4.35

Martin, G., Gollan, P. J., & Grigg, K. (2011). Is There a Bigger and Better Future for Employer Branding? Facing Up to Innovation, Corporate Reputations and Wicked Problems in SHRM. *International Journal Of Human Resource Management*, *22*(17), 3618–3637. https://doi.org/10.1080/09585192.2011.560880

Mayring, P. (2010). Qualitative Inhaltsanalyse. *In VS Verlag für Sozialwissenschaften eBooks*, 601–613. https://doi.org/10.1007/978-3-531-92052-8_42

McKinsey & Company. (2023). The Beauty Market in 2023: A Special State of Fashion Report. *McKinsey & Company*. https://www.mckinsey.com/industries/retail/our-insights/the-beauty-market-in-2023-a-special-state-of-fashion-report#/

Mears, A. (2014). Aesthetic Labour for the Sociologies of Work, Gender, and Beauty. *Sociology Compass*, *8*(12), 1330–1343. https://doi.org/10.1111/soc4.12211

Peiss, K. (2000). On Beauty . . . and the History of Business. *Enterprise & Society*, *1*(3), 485–506. https://doi.org/10.1093/es/1.3.485

Pounders, K. R., Babin, B. J., & Close, A. G. (2014). All the Same to Me: Outcomes of Aesthetic Labour Performed by Frontline Service Providers. *Journal of the Academy of Marketing Science*, *43*(6), 670–693. https://doi.org/10.1007/s11747-014-0407-4

Rafiq, M., & Ahmed, P. K. (2000). Advances in the Internal Marketing Concept: Definition, Synthesis and Extension. *Journal of Services Mar-*

keting/the Journal of Services Marketing, *14*(6), 449–462. https://doi.org/10.1108/08876040010347589

Reis, G. G., Lacombe, B. M. B., & Trullén, J. (2017). Workplace Authenticity as an Attribute of Employer Attractiveness. *Personnel Review*, *46*(8), 1962–1976. https://doi.org/10.1108/pr-07-2016-0156

Ružić, E., & Benazić, D. (2023). Dimensions of Attractiveness in Employer Branding and the Value Proposition Framework for Young Employees. *Ekonomski Vjesnik*, *36*(1), 89–100. https://doi.org/10.51680/ev.36.1.7

Santos, R. C. D., Silva, M. J. D. B., Costa, M. F. D., & Batista, K. (2023). Go Vegan! Digital Influence and Social Media Use in the Purchase Intention of Vegan Products in the Cosmetics Industry. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01034-7

Sivertzen, A., Nilsen, E. R., & Olafsen, A. H. (2013). Employer Branding: Employer Attractiveness and the Use of Social Media. *Journal of Product & Brand Management*, *22*(7), 473–483. https://doi.org/10.1108/jpbm-09-2013-0393

Soprano, M., Roitero, K., Barbera, D. L., Ceolin, D., Spina, D., Demartini, G., & Mizzaro, S. (2024). Cognitive Biases in Fact-Checking and Their Countermeasures: A Review. *Information Processing & Management*, *61*(3), 103672. https://doi.org/10.1016/j.ipm.2024.103672

Talamas, S., Mavor, K. I., & Perrett, D. I. (2016). Blinded by Beauty: Attractiveness Bias and Accurate Perceptions of Academic Performance. *PloS One*, *11*(2), e0148284. https://doi.org/10.1371/journal.pone.0148284

Theurer, C. P., Schäpers, P., Tumasjan, A., Welpe, I. M., & Lievens, F. (2021). What You See Is What You Get? Measuring Companies' Projected Employer Image Attributes via Companies' Employment Webpages. *Human Resource Management*, *61*(5), 543–561. https://doi.org/10.1002/hrm.22085

Vonk, L. (2020). Peer Feedback in Aesthetic Labour: Forms, Logics and Responses. *Cultural Sociology*, *15*(2), 213–232. https://doi.org/10.1177/1749975520962368

Wallace, M., Lings, I., & Cameron, R. (2012). Industry Branding: Attracting Talent to Weaker Profile Industries. *Asia Pacific Journal of Human Resources*, *50*(4), 483–502. https://doi.org/10.1111/j.1744-7941.2012.00040.x

Warhurst, C., & Nickson, D. (2007a). A New Labour Aristocracy? Aesthetic Labour and Routine Interactive Service. *Work, Employment and Society*, *21*(4), 785–798. https://doi.org/10.1177/0950017007082887

Warhurst, C., & Nickson, D. (2007b). Employee Experience of Aesthetic Labour in Retail and Hospitality. *Work, Employment And Society*, *21*(1), 103–120. https://doi.org/10.1177/0950017007073622

Warhurst, C., Nickson, D., Witz, A., & Cullen, A. M. (2000). Aesthetic Labour in Interactive Service Work: Some Case Study Evidence from the "New" Glasgow. *Service Industries Journal/the Service Industries Journal*, *20*(3), 1–18. https://doi.org/10.1080/02642060000000029

List of Abbreviations

AI Artificial Intelligence

AL Aesthetic Labour

CSR Corporate Social Responsibility

EB Employer Branding

EmpAt Employer Attractiveness

FMCG Fast-Moving-Consumer-Goods

Gen Z Generation Z

HR Human Resources

HRM Human Resource Management

SHRM Strategic Human Resource Management

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Understanding Participation Trends in the Avani 5G Patent Pool: A Descriptive Analysis

Justus Scharfstädt

Technical University of Munich

Abstract

Patent pools play a critical role in adapting to new technology standards by simplifying the licensing of Standard-Essential Patents (SEPs). The Avanci 5G pool provides a licensing platform for vehicle manufacturers to access essential technologies for connected vehicles. Despite the growing importance of these pools, there is limited research on the dynamics of participation and its implications for different stakeholders. This study provides a descriptive analysis of the Avanci 5G pool, particularly focusing on the development of the SEP and licensee coverage since its inception. It uses a mixed-methods approach and combines multiple data sources for the coverage estimates. My findings reveal substantial growth in both SEP and licensee coverage but highlight a significant gap in the participation of Asian automakers, potentially due to perceived anti-competitive risks. Policy makers and relevant stakeholders should aim to foster a competitive and innovative environment in SEP licensing.

Keywords: patent pool; licensing; connected vehicles; innovation; anti-competitive risks

1 Introduction

Since the 1980s, mobile communications have rapidly gained importance and increasing data demands have driven significant technological advancements Khanh et al. (2022, pp. 1–3). From 1G to 5G, the technology has evolved through five generations.

Large-scale 5G adoption began in 2019, addressing the limitations of previous generations by offering higher data rates, reduced latency, and enhanced connectivity to support the growing Internet of Things (IoT) (Rabbi, 2020). A report published by Statista (2024b) illustrates the remarkable IoT development, with a forecast of more than a doubling from 18 billion in 2024 to 39.6 billion connected IoT devices in 2033. This connectivity enables several *"breakthrough"* applications including smart cities, smart agriculture, and smart retail Khanh et al. (2022, p. 1).

However, the automotive industry is one that is experiencing a significant impact from the evolution of 5G Lanctot et al. (2023, pp. 4–16). An increase from 450 million connected vehicles in 2025 to over 835 million in 2035 is expected (Statista, 2022). Consequently, the share of connected vehicles will increase from 66% in 2023 to 96% in 2030 (Counterpoint, 2024b; Statista, 2020). A report published by Precedence Research (2024) illustrates that the connected car market is mainly located in Asia Pacific (39%), followed by North America (28%), Europe (25%), Latin America and Africa.

The shift from 4G to 5G is seen as a transformation that enhances vehicle connectivity, safety, and infotainment. This is crucial to improve features like autonomous driving and remote vehicle diagnostics. As Kasim Alfalahi, CEO of Avanci, states, *"5G is no longer the next big thing in the automotive space"* Lanctot et al. (2023, p. 3). As a result, 5G car revenues are expected to exceed non-5G car revenues by the latest of 2028 Lanctot et al. (2023, p. 5). According to Counterpoint (2024a), China will lead the connected car market, followed by the United States of America (USA), Japan and Germany.

While vehicles become more connected, the automotive industry is dependent on technology standards that ensure compatibility and functionality across diverse systems (Pohlmann, 2017). These standards are often based on patented technologies, which are known as SEPs if they are necessary for the implementation of the standard Baron and Pohlmann (2018, p. 505). Therefore, Original Equipment Manufacturers (OEMs)[1] need to navigate a complex patent licensing landscape and face potentially high licensing costs (Pohlmann, 2017).

Patent pools, which aggregate SEPs from multiple patent holders, play an important role in simplifying the licensing process by reducing transaction costs and minimizing litigation risks (Lavenue et al., 2023). Since 2016, Avanci acts as a pioneering marketplace for multiple generations of technology patents within the automotive industry (Avanci, 2024h). Building on the success of the 4G pool, Avanci launched their 5G vehicle pool on August 16, 2023, and since then attracted many licensors and licensees (Avanci, 2024f).

Nonetheless, there have been notable developments along SEPs and patent pools in the automotive industry.

[1] In this context, I refer to end-use vehicle manufacturers as OEMs.

DOI: https://doi.org/10.5282/jums/v11i1pp164-180

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

First, the European Commission (2023b) published a proposal on SEP regulations to encourage innovation and Fair, Reasonable and Non-Discriminatory (FRAND) access to technology. Second, although patent pools are beneficial to the licensing process, there is ongoing resistance over anti-competitive practices and lack of transparency (ACEA, CLEPA, CEMA, CECE, EGMF, & FEM, 2024; State Administration for Market Regulation, 2024). Third, the creation of counter pools by OEMs highlights their willingness to compete by combining negotiation forces on the licensee side (Bundeskartellamt, 2024).

Given these developments, it is important to elaborate on how emerging technologies and licensing practices are shaping the future of automotive innovation. Despite extensive research, gaps remain in understanding the impact of the participation dynamics in recently established patent pools like the Avanci 5G pool. This thesis seeks to address these gaps by investigating the evolution of SEP licensing and its effects on the automotive industry. Specifically, it analyzes participation trends in the Avanci 5G pool and provides estimates of SEP and licensee coverage developments.

This thesis is organized as follows: Section 2 reviews the literature on standardization, licensing strategies and the role of patent pools. It also derives the research question. Section 3 explains the research methodology and the use of different data sources. After that, Section 4 presents and visualizes the results. Section 5 evaluates the implications of the findings and relates them to the literature to discuss key differences and similarities. Finally, Section 6 summarizes the thesis with some concluding comments.

2 Literature Review and Motivation of Research

A considerable number of studies and research papers covers standardization, patent pools and the overall SEP licensing ecosystem. In this section, I review these terms and look at how to apply them to the automotive industry. Afterwards, I illustrate the derivation of my research question.

2.1 Standardization and Standard-Essential Patents

In recent years, the relationship between technology standards and patents has attracted significant interest from academics and policymakers Baron and Pohlmann (2018, pp. 504–534).

The European Telecommunications Standards Institute (ETSI) defines a standard as a document, created through agreements and approved by a recognized body, that provides rules and guidelines to achieve order in a given context ETSI (2024c, p. 14).

Several industries are dependent on advanced technologies and therefore heavily rely on standardization. In addition, many standards are based on cutting-edge patented technologies, resulting in a large number of innovations protected by SEPs Baron and Schmidt (2019, p. 5).

This important interrelation between standards and SEPs is exemplified by technology-intensive sectors like the information and communications technology industry. Different networks, from 1G to 5G, enable the development of hyperconnectivity in the IoT, which is crucial for this thesis and the examination of SEP licensing in the automotive industry (European Commission, 2024b).

The literature on standardization identifies three primary categories of standards: de facto, formal and de jure (Bellflamme 2002, pp. 153–176; Chiesa et al. 2002, pp. 431–450; David and Greenstein 1990, pp. 3–41). De facto standards emerge through widespread market adoption and become dominant through popularity or market share. Formal standards are developed and approved by recognized Standard-Setting Organizations (SSOs), while de jure standards are mandated by legislative bodies.

Standards are often developed by SSOs that require their members to license the patented technology that is necessary for the implementation of a specific standard Baron and Pohlmann (2018, pp. 504–534). The main functions of SSOs are discovery, standardization, and regulation Lerner and Tirole (2015, pp. 547–586). Discovery includes understanding and verifying the value of the technology. Shifting market expectations towards the discovered technology is done in the standardization function. With its third function, the regulation, SSOs make sure that SEP holders grant licenses on FRAND terms.

FRAND conditions are a fundamental property of protecting technology implementors with fair royalty rates but also giving an appropriate reward for research and development (R&D) investments and technology availability to the patent owner Lemley and Shapiro (2013, p. 1137). Calculating "fair and reasonable" rates and defining "non-discriminatory" conditions is *"inherently vague"* Geradin (2020, pp. 25–38). The comparable license approach, which analyzes the rates of other similar licensing agreements, is widely used to determine the royalty rate. Alternatively, the top-down approach first calculates the aggregate royalty for all SEPs covering a specific standard, and then distributes this among individual SEPs and their holders Bekkers et al. (2022, p. 5). "Non-discriminatory" means that any company that wants to implement a technology should have access to the respective license and pay the same royalty rate as others.

During the standard development process, SSOs iteratively work in interaction with manufacturers and incorporate various stakeholders' feedback to ensure comprehensive and effective standards Baron and Pohlmann (2018, pp. 462–503).

In addition to other advantages, standards also ensure safety, reliability, and protect the interests of businesses and users (ETSI, 2024c). Furthermore, standardization shapes the global economy by establishing quality benchmarks, solving connectivity problems, and enabling technological advancement which result in increasing innovation and customer satisfaction Kanevskaia (2020, pp. 161–178).

Despite these advantages, standards also face various challenges and limitations. Verifying FRAND conditions and determining whether patents are truly essential are major issues. Because non-disclosed SEPs can be legally unenforceable and no penalties exist for disclosing non-SEPs,

"over-declaration" can occur. According to a study by the European Commission (2023a), it is estimated that only 20-30% of declared SEPs are truly essential, with even lower tendencies of 10-20% for 5G.

Additional concerns arise as granted SEPs might give the holder a temporary monopoly due to individual price setting and exclusion of implementors that do not agree on licensing conditions Blind and Thumm (2004, pp. 1583–1598). Three reasons for welfare losses related to standards are identified in Barnett (2018, pp. 1321–1324). First, "patent thickets" occur when overlapping patents must be licensed from multiple owners. Second, the situation where numerous patent holders each demand high royalties for their contributions to a single standard is known as "royalty stacking". Third, "patent holdup" happens when a patent owner demands excessive royalties after a company has already made a significant investment in the technology.

Regarding the SSOs, different approaches for standard setting, including policies on membership, participation, and Intellectual Property Rights (IPRs) management might lead to inconsistency and conflicts with the approaches of other SSOs Baron and Pohlmann (2018, pp. 462–503).

Overall, while *"standardization is a key contributor to industrial innovation and competitiveness"* European Commission (2023b, p. 1), it also faces challenges that require careful management and regulation to ensure FRAND access to essential technologies.

To address these issues, the European Commission (2023b) proposed a regulatory framework for SEPs in April 2023. They seek to benefit European Union (EU) consumers and businesses by promoting the use of standardized technologies, fostering innovation within the EU, and encouraging competitive participation in EU and global markets. The feedback on this issue shows a mixed perception among the stakeholders (European Commission, 2024a). Several licensors and research groups are concerned about slowing down the negotiation process, reducing R&D investment and having a negative impact on international competitiveness. On the other hand, consumer groups and licensees, including Mercedes-Benz (Muller, 2023), support the proposal to ensure transparency, fairness and a more balanced SEP ecosystem.

In a completely different approach, the United Kingdom (UK) Intellectual Property Office launched a SEPs Resource Hub[2] that gathers SEP-related information and aims to assist UK businesses in navigating the complex licensing landscape (UK Government, 2024). It is the first of its kind and has been developed in collaboration with industry stakeholders (Brazier, 2024).

2.2 Types of Licensors

The licensors, or patent owners play a crucial role in the licensing ecosystem. This section provides a detailed examination of the different categories of licensors and highlights their functions and strategies within the market.

[2] UK SEPs Resource Hub

In this context, Dewatripont and Legros (2013, p. 928) distinguish between upstream and vertically-integrated firms. First, upstream firms operate at the initial stages of the production process and develop technologies. Licensing revenues often cover their R&D costs and therefore increase profits. Their business model incentivizes them to engage in patent *"padding"*, i.e. increasing the number of patents, even if some are not essential. Second, vertically integrated firms develop technologies and generate licensing revenues (upstream) but also utilize external licenses to manufacture goods (downstream). This leads to a more balanced motivation to declare patents.

In addition to generating revenues through licensing patented technologies, patent owners have the option to sell their IPRs. They may choose to divest to other licensors, as exemplified by the patent assignment from OPPO to ZTE (Müller, 2024b). Such transactions are often strategic decisions to resolve litigation disputes and optimize their portfolios.

Furthermore, patents can also be sold to Non-Practicing Entities (NPEs) that do not engage in own development but strategically acquire IPRs to enforce and license them Schwartz and Kesan (2014, pp. 426–456). By that, NPEs can also mitigate the risk of high litigation costs for smaller inventors Cohen et al. (2018, p. 2). However, the presence of NPEs introduces a layer of complexity in patent licensing and NPE transferring. Especially the aggressive *"you pay immediately or we go to court"* strategy European Commission (2014, p. 67) leads to rising defense costs. Data from the European Commission (2014, p. 68) indicate that the activities of NPEs, specifically SEP transfers, have increased significantly since 2005.

2.3 Types of Licensees

On the opposite side of the licensing agreement are the licensees. This section examines the distinction between licensing to end-product manufacturers and component suppliers.

In a multi-layered supply chain, patent owners can primarily license their technology to component suppliers or end-product manufacturers. Geradin and Katsifis (2021, p. 8) evaluate the debate on whether SEP holders must license to all implementers, including component suppliers, or can choose to license only at the end-product level. On the one hand, supporters of supplier licensing argue that it can potentially lower transaction costs, prevent over-compensating SEP holders with excessively high royalty rates, and reduce legal uncertainty for component manufacturers operating without a license. On the other hand, proponents of licensing SEPs only to end-product manufacturers state that this approach aligns with traditional industry practices and ensures fair compensation for the patent owners. Nonetheless, Martinez (2019, p. 649) stresses that component suppliers have low profit margins and are therefore unable to pay *"adequate"* royalties to SEP holders.

The Nokia vs. Daimler case serves as a pertinent example that illustrates the licensee differentiation problem (Raturi, 2021). In 2019, Nokia sued Daimler for using telematic control units without a license, but Daimler, being an end-product manufacturer, argued that its suppliers should obtain the licenses. In August 2020, the Mannheim Regional Court decided that Daimler did not show willingness to obtain a license and granted Nokia injunctive relief. Following the decision, the Bundeskartellamt suggested that the court should refer competition concerns to the European Court of Justice to address whether SEP-holders can choose which supply chain entities to license. However, the Mannheim Court refused this (Klos, 2020; Raturi, 2021).

Several challenges arise from the refusal to license SEPs to component suppliers. The competition rules published by the European Commission (2017) state that licensing a SEP under FRAND conditions means that it *"should be available to any potential user of the standard"*. Furthermore, SEP holders might reduce competition by withholding licenses that are necessary for manufacturing and selling standard-compliant products Nazzini (2017, p. 21).

Examples for licensing only to end-product manufacturers and refusing to license component suppliers exist, namely the Avanci patent pools (Muller, 2022). In 2019, the automotive component supplier Continental sued Avanci for this behavior Koh (2019, p. 2).[3] In contrast, the Via Licensing Alliance (2024) follows another approach and offers licenses to manufacturers, suppliers, and other potential users.

Despite the decision on whom to license to, there is another important development around licensees, particularly connected vehicle manufacturers. In February 2024, Toyota, previously only a licensee, acquired multiple SEPs from the electronics manufacturer and SEP licensor OPPO (Müller, 2024d; Shan, 2024). With deals like this, licensees aim to solve the "patent dilemma", which refers to the disadvantages they face in SEP litigation against licensors. Benefits of such deals include the possibility to counter sue, having a broader portfolio, and using acquired patents for licensing agreements to lower own licensing costs. However, this also adds further complexity to the licensing ecosystem.

2.4 The Role of Patent Pools

To bridge the gap between licensees and licensors, patent pools serve as a critical mechanism. In the following, I analyze their objectives, features and development.

Literature shows that licensing negotiations are expected to become more complex, which makes it important for patent owners to provide a variety of licensing mechanisms Nikolic and Galli (2021, pp. 1–8). Licenses can be obtained through bilateral agreements between licensees and licensors, which are often resource-intensive and time-consuming, or via a joint licensing arrangement, such as a patent pool SEPs Expert Group (2021, p. 27).

Dequiedt and Versaevel (2013, p. 59) define patent pool as *"cooperative agreements between two or more firms to license their related patents as a bundle"*. To access this bundle, a royalty rate needs to be paid to the pool, which is then divided among the licensors. Thereby, patent pools enable a streamlined licensing process with reduced transaction costs and litigation uncertainties as well as predictable licensing costs. The "one-stop shopping" feature highlights efficiency and access advantages for many market participants.

That said, the presence of patent pools can also result in negative effects on competition and innovation SEPs Expert Group (2021, p. 27). There is a risk of anti-competitive practices, such as price-fixing, market fragmentation or excluding potential users. Patent pools therefore have to comply with several conditions: the patents included need to be essential and verified by independent third parties, members agree to follow FRAND conditions, and licensors must keep the right to license their patents independently.

Historically, patent pools have been the earliest form of patent aggregation that first emerged in the USA in the late 19th century Nikolic and Galli (2022, p. 4). After initially facing antitrust problems in mid-1900s, they re-emerged in the 1980s for audio-visual technologies. Regarding IoT networks, a study by Pohlmann (2015) found that patent pools were not common for technology standards. In 2015, only 9% of globally declared SEPs were licensed through pools.

This low adoption can be explained by three factors. First, licensors do not want to lose leverage when negotiating cross-licensing agreements. Second, the maximum price of the pool might be below the one favored by the patent owner. Lastly, the complexity and high setup costs make it difficult to agree on conditions with all stakeholders of a patent pool (Nikolic and Galli 2022, p. 4; Pohlmann 2015).

However, multiple key developments mark the evolution of patent pools. *"Major signals"* for increasing importance and industry coverage of patent pools are analyzed by Mossoff (2023). Over the past two decades, patent pools have emerged in numerous innovative industries, including eye surgery, audio technology, smart TVs, 4G and 5G, among many others. The significance is underlined by the success of the Avanci 4G vehicle pool, which covers over 80% of connected cars (Müller, 2024c), and the merger of Via Licensing and MPEG LA to form the Via Licensing Alliance (Via Licensing Alliance, 2023), the largest patent pool in history.

Although this trend is evident, resistance partially remains. Several parties criticize patent pools and raise concerns on lacking FRAND conditions, non-transparent information, and call for monopoly risk assessments related to pools' licensing practices. This is exemplified by two recent letters[4] addressed to the Avanci 5G vehicle pool (ACEA, CLEPA, CEMA, CECE, EGMF, & FEM, 2024; State

[3] Continental, the plaintiff, argued that Avanci's unfair approach aims to uphold the high royalty rates that only vehicle manufacturers with higher profit margins can afford.

[4] Six automotive groups express the need of transparency about essentiality verification and the royalty rates, and the Chinese State Administration for Market Regulation reminds of the risk of monopolistic licensing activities.

Administration for Market Regulation, 2024). After the US Department of Justice ruled that Avanci's pool are unlikely to harm competition and the EU took steps to enhance transparency, China is now challenging the pool's activities (Chun, 2024).

Additionally, a recent development goes beyond the written resistance: the approval for a licensing negotiation group (LNG)[5], which plans to lower negotiation efforts and transaction costs by combining the strength of multiple licensees (Bundeskartellamt, 2024). Despite this, patent owners fear that counter pools could grant too much power to the licensees. Although the formation is approved, the Bundeskartellamt (2024) emphasizes several conditions for LNGs: voluntary participation, restricted activities, and a maximum 15% share of the purchasing market that must not be exceeded. Schüll (2024), patent attorney and partner at Cohausz & Florack, says that the decision of the Bundeskartellamt proves that *"Germany and the EU play a leading role in the field of SEPs"*.

This development follows other relevant agreements in the licensing landscape. Patent pools already agreed with LNG-like companies that operate as patent aggregators, as exemplified by two licensing agreements of Sisvel and RPX (Dini, 2024).[6] The collaborations of patent pools and patent aggregators demonstrate the efficiencies of working together in terms of reduced transaction costs and balanced interests of patent owners and technology implementors (Houldsworth, 2024). Therefore, other licensees may be incentivized to form LNGs as well.

2.5 Derivation of the Research Question

The literature review underscores the increasing significance of SEPs and patent pools in the licensing ecosystem. The integration of technology standards, particularly SEPs, involves navigating complex challenges.

On one hand, patent pools might be a potential solution to some of these challenges by streamlining licensing processes, reducing transaction costs, and mitigating litigation risks. On the other hand, the successful establishment of patent pools depends on multiple factors including attracting new members. Recent developments in the licensing landscape reveal ongoing resistance, critical concerns, and the formation of counter pools, the implications of which remain unclear. Further, different regulatory frameworks are proposed.

As patent pools continue to evolve rapidly, it becomes essential to evaluate the participation trends and determine the pool's coverage of SEPs and licensees. There is a literature gap regarding the analysis of participation trends in newly established licensing models such as the Avanci 5G vehicle pool. Therefore, this thesis aims to fill the existing gap by providing valuable and detailed insights into the emerging dynamics within the Avanci 5G pool. It discusses broader implications for OEMs, patent owners, and regulators. As a result, I formulate the following research question:

> **How have the patent and licensee coverage evolved within the Avanci 5G pool since its inception?**

3 Research Method

This section outlines the research methodology employed in this study, including the utilization of different data sources for the Avanci 5G pool, its members, SEP and licensee coverage calculations, revenue estimations and related court cases. The analysis covers the period from the launch of the Avanci 5G pool on August 16, 2023, (Avanci, 2024b) until August 31, 2024.

3.1 Presentation of the Avanci 5G Pool

Beginning with Avanci, which positions itself as *"the independent global leader in joint licensing solutions"* (Avanci, 2024e), the pool administrator offers patent pools across various industries, including vehicles, IoT and broadcast (Avanci, 2024h). This thesis focuses on the vehicle sector, particularly the 5G pool (Avanci, 2024c).

Compared to the 4G pool, the 5G pool covers SEPs for 2G, 3G, 4G and 5G technology (Avanci, 2024f). This includes the Vehicle-to-Everything connectivity that enables communication to other vehicles, pedestrians, and the overall infrastructure. The Avanci 5G pool states that it features *"market-driven pricing"*, which can be divided into two different royalty rates. According to an interview with Avanci vehicle president Laurie Fitzgerald (Jiang, 2024), the pricing structure was established based on feedback from multiple stakeholders. The $29 per vehicle *"early-licensee pricing"* applies to all licensees who sign the license agreement before February 16, 2024, or before their first 5G connected vehicle sale. Depending on the case, "*other conditions apply*" (Avanci, 2024c). The *"base running royalty"* amounts to $32 per vehicle, resulting in a 60% increase compared to the $20 of the 4G pool (Avanci, 2024a). Avanci (2024f) also states that the prices might change in the future.

The Avanci 5G pool seeks to transform shared technology by securing licensors' SEPs and providing licensees with a broad patent coverage, both optional and on FRAND terms (Avanci, 2024f). Consequently, it aims to increase efficiency in terms of reduced negotiation time and transaction costs.

3.2 Data Collection of Avanci Members

To create a foundation for the descriptive analysis of the Avanci 5G pool, it is essential to determine the joining dates of the participating licensors and licensees.

There are multiple ways to collect the joining dates. I start by monitoring the members' websites and social media

[5] The "Automotive Licensing Negotiation Group" was requested by Volkswagen, Mercedes, BMW and ThyssenKrupp.

[6] Sisvel, owner of multiple patent pools, and RPX, a patent aggregator for approximately 320 technology implementors, made licensing agreements for WIFI and Video SEPs.

statements, for example on LinkedIn[7]. Additionally, Avanci announces new entrants on their "News & Views" page[8] and displays the corresponding logos on their website. Industry experts and blogs related to SEPs and the licensing ecosystem also post regular updates on the Avanci 5G pool. For instance, Florian Müller, author of the blog ip fray[9] (formerly Foss Patents[10]), provides *"in-depth reporting and analytical commentary on intellectual property disputes and debates"* (Müller, 2024a). IAM Media[11] exemplifies another source with news articles on patent updates within the automotive industry.

However, these sources do not capture all joining members. To address this limitation, I employ the Internet Archive Wayback Machine[12], following the approach of Baron and Ménière (2014, p. 55) who evaluate descriptive statistics for patent pools. This approach enables me to track the joining dates approximately by comparing the member logos across different versions of the Avanci 5G pool website. Given that not all logos are consistently published on the website, my data collection is based on a combination of the mentioned sources. Appendix 1 lists all participating licensors and licensees with their entry dates.

For each licensee, I further assess whether they joined before or after the royalty rate increase on August 16, 2024. However, it is beyond the scope of this thesis to track the timing of their first 5G-connected vehicle sale to determine whether those who joined after the price increase actually pay the higher rate. Nonetheless, I examine two other company-specific characteristics of the licensees: their headquarter locations and vehicle segment.

3.3 Collection of ETSI Data

To obtain the necessary data to address the research question concerning the SEP coverage of the Avanci 5G pool, I need to assess the share of SEPs held by Avanci licensors compared to the total number of SEPs declared.

The starting point for the analysis is set with an extract of the ETSI IPR database that is released twice a year in the special report SR 000 314 (ETSI, 2024b). In its report, ETSI (2024a, p. 4) mentions that it does not verify the relevance or essentiality of declared patents or patent applications. Moreover, the SSO does not conduct any research on other patents that might be or become essential. However, if ETSI members become aware of any essential IPR, they are obliged to make *"reasonable efforts"* to notify ETSI.

I use the version V2.35.1 of the ETSI special report, that was published in April 2024 and contains 1,044,211 entries in its raw format. To identify all unique patents disclosed to ETSI, I follow an approach similar to the one used by Bekkers et al. (2020). In the first step, I keep only the relevant entries by cleaning the raw data using the following fields:

1. **'COMP_LEGAL_NAME'** is the *"legal name provided by the declarer to identify the declarant company"* ETSI (2024a, p. 14).
2. The columns **'2G, 3G, 4G, and 5G'** indicate whether *"any of the disclosed standards/work items/specifications disclosed 3GPP technologies correspond to '2G', '3G', '4G', or '5G'"* ETSI (2024a, p. 16).
3. **'DECL_IS_PROP_FLAG'** shows *"whether the declarant and/or its affiliates declared to be the proprietor of the IPR(s) disclosed in the attached IPR Information Statement Annex, or not"* ETSI (2024a, p. 14).
4. **'Ess_To_Standard'** indicates *"whether the patent is disclosed as essential to a specified standard/work item/specification"* ETSI (2024a, p. 15).
5. **'PATT_ID'** is a *"unique numeric identifier of a patent in the IPR database"* ETSI (2024a, p. 15).
6. **PATT_APPLICATION_NUMBER** identifies the *"patent application number of the patent identified by PATT_ID"* ETSI (2024a, p. 15).
7. **PUBL_NUMBER** is the *"patent publication number of the patent identified by PATT_ID"* ETSI (2024a, p. 15).
8. **'IPRD_SIGNATURE_DATE'** indicates the *"date of the IPR declaration submission to ETSI"* ETSI (2024a, p. 14).
9. **'DIPG_PATF_ID'** is the *"unique numeric identifier of a disclosure's ETSI patent family in the IPR database"* ETSI (2024a, p. 15).

Afterwards, I add the field **'COMP_NAME_CONSOLIDATED'** to display the correct consolidated names for the relevant applicant companies. Table 1 shows the cleaning process at patent family level, including the additional step 11, and the remaining entries after the different steps.

Bekkers et al. (2020, p. 11-14) eliminate duplicates with the same PATT_APPLICATION_NUMBER, PUBL_NUMBER and DIPG_PATF_ID. After this step, some duplicates with identical PATT_IDs remain. Therefore, they try to identify all relevant ETSI patents in combination with the PATSTAT[13] database released by the European Patent Office. Their procedure with its materials and proficiency exceeds the scope of this thesis. My methodology includes several other steps to reach a useful approximation.

I start my data cleaning process by eliminating all entries without a COMP_LEGAL_NAME, as these entries are not associated to any legal entity. For the focus of my research, I further ensure that only patents corresponding to 2G, 3G, 4G or 5G (identified by a '1' entry) remain. After that, setting DECL_IS_PROP_FLAG to true ensures that the respective declarants are also the proprietors of the IPR. In the

[7] LinkedIn
[8] Avanci "News & Views"
[9] ip fray
[10] FOSS Patents
[11] IAM Media
[12] Wayback Machine
[13] European Patent Office PATSTAT

Table 1: ETSI data cleaning process at patent family level

Steps	Patent family level	# Entries
0	Raw	1,044,211
1	Remove (blank) COMP_LEGAL_NAME	1,044,111
2	Remove (blank) 2G-5G	943,868
3	Set DECL_IS_PROP_FLAG = TRUE	943,809
4	Set Ess_To_Standard = TRUE	941,632
5	Remove "" OR "-" PATT_APPLICATION_NUMBER	919,135
6	Remove (blank) OR "" OR "n/a" OR "-" OR "pending" PUBL_NUMBER	857,187
7	Remove duplicates using PATT_ID	224,877
8	Remove duplicates using PATT_APPLICATION_NUMBER	216,750
9	Remove duplicates using PUBL_NUMBER	184,059
10	Sort IPRD_SIGNATURE_DATE ascending	184,059
11	Remove duplicates using DIPG_PATF_ID (- 1 blank)	75,528
12	Add COMP_NAME_CONSOLIDATED with consolidated licensor name	75,528

fourth step, I remove all patents that are disclosed as non-essential by setting Ess_To_Standard to true. Furthermore, there are several irrelevant entries in the PUBL_NUMBER or PATT_APPLICATION_NUMBER fields that contain missing or unusual values (e.g. “na”, “pending” or “-”) that I delete as well. After eliminating irrelevant entries, I focus on duplicates. The report published by ETSI (2024a, p. 15) states that the PATT_APPLICATION_NUMBER and PUBL_NUMBER are identified by the PATT_ID. Nonetheless, I believe that it would be insufficient to eliminate entries based only on the PATT_ID, as duplicates with identical values in the other two fields may remain because of database updates or corrections. Thus, I check for duplicates in all three fields individually. To retrieve the remaining records with the oldest disclosure dates, I order them in ascending order based on the IPRD_SIGNATURE_DATE. This is also done by Bekkers et al. (2020, p. 13). I perform an additional step for the analysis at patent family level: duplicates are further eliminated using the DIPG_PATF_ID.

All the steps lead to a remaining set of 184,059 entries at patent- and 75,528 at patent family level, significantly less than the initial 1,044,211.

Lastly, I ensure that all licensors are accurately listed on company level by adding the field COMP_NAME_CONSOLIDATED. For instance, I aggregate “NOKIA MOBILE PHONES” and “Nokia Corporation” to “Nokia”. In this context, it is necessary to note that for some NPEs, information on the number of patents they hold are unavailable.[14] Consequently, I transfer the patents to the NPE only if online sources provide proof of ownership. Sterzi et al. (2020, p. 4) also highlight this problem of limited transparency of patent ownership and mention that NPEs are *“gaming the system”*, which *“poses serious challenges to competition authorities and regulators”*. Appendix 2 shows the licensor consolidation and monthly SEP analysis at patent level.

3.4 Collection of Judicial Data

Court cases and judicial rulings significantly impact the licensing ecosystem, specifically negotiation strategies and patent pool participation trends OECD (2019, pp. 24–36).

To get an overview of ongoing court cases and recent judgements regarding automotive SEP licensing and the Avanci 5G pool, I review multiple online sources. As already stated in the Data Collection of Avanci Members (see section 3.2.), ip fray[15] (formerly FOSS Patents[16]) regularly posts detailed updates on relevant judicial data.

For this thesis, I begin by looking for cases where Avanci licensors are plaintiffs and potential licensees are defendants. However, there is not a single 5G infringement case between an Avanci licensor and a vehicle manufacturer (Müller, 2024c). Therefore, I also analyze cases with Avanci licensees acting as defendants against potential licensors. Furthermore, judicial rulings involving Tesla are also relevant for my research because the automaker holds an Avanci 4G license and tries to get a preferential 5G rate (Müller, 2024c).

[14] Applies for: 5G IP Holdings, China Mobile, China Telecom, China Unicom, Coolpad, Crystal Clear Codec, Emergent Mobile, EQUO Innovative Partner, Harfang IP, Intellectual Discovery, IP Innovation, JVCKENWOOD, KDDI, MiiCs, New Radio, Nippon Telegraph and Telephone, Offino, Oswego, Technology in Ariscale, TNO, Top Quality Telephony, UU Ventures.

[15] ip fray

[16] FOSS Patents

3.5 Data Consolidation and Processing

Following the collection of all relevant information and data needed for the descriptive analysis, I will consolidate these to get an overview, extract key findings, and visualize them in form of figures and tables.

The participation trends of the licensors and licensees provide a good starting point. Building on that, I use the proportion of participants joining at each date for a time-to-license calculation.

For the first part of the research question, the SEP coverage estimation, it is essential to combine the licensor and ETSI data. I collect the amount of SEPs covered by comparing the participating licensors' joining dates (see section 3.2) with the IPRD_SIGNATURE_DATE and their name with the COMP_NAME_CONSOLIDATED. The Avanci share is then calculated by dividing the amount held by the members by the total number of declared SEPs.[17] I perform this calculation at patent- and patent family- as well as monthly- and quarterly level, resulting in 13 and 4 observations per licensor, respectively. However, it is important to note that my results could differ from the actual SEP coverage of the Avanci 5G pool.

For the second part of the research question regarding the licensee coverage estimation, I begin by compiling a list of OEMs worldwide, ranked by revenue. Although Avanci's licensees produce passenger, industrial, agricultural and defense vehicles[18], this thesis targets only passenger vehicles. I create a python web scraper that uses the list of 147 automakers published by Statista Passenger Cars Worldwide[19], which includes their headquarters locations and revenue forecasts for 2024. I also manually add Polestar and VinFast to the list, as they are not included.[20] All passenger car manufacturers with a forecast of less than USD 10 million are removed to exclude unnecessary players, leaving 129 OEMs. To estimate the licensee coverage of the Avanci 5G pool, I identify which OEMs from the list hold a license. This results in two coverage estimates: geographic by continent and financial by revenue. While the geographic coverage is analyzed as of August 31, 2024, the revenue coverage development is evaluated monthly from August 31, 2023, to August 31, 2024.

It is crucial to recognize that it would be inaccurate to assume that all revenues will be generated through connected cars. The share of revenues from connected cars can vary significantly depending on the company's focus. Given the scope of this thesis, I follow the estimation used by Counterpoint (2024b) and apply a fixed amount of 66% for all OEMs. To calculate the coverage, I divide the cumulative projected connected car revenues of the OEMs participating in Avanci by the total forecasted connected car revenues for the entire automotive market.

I also apply the 66% share for another insightful estimation: calculating the total revenue generated by the Avanci 5G pool in 2023 and 2024.[21] I collect the 2023 vehicle sales for all licensees using the information from their websites and press releases.[22] For 2024, I forecast an average increase in connected car sales of 2.3%, which is also predicted by Statista (2024a) for the overall automotive market. To estimate the licensees' payments, I multiply their connected vehicle sales (66% of their total unit sales) by the royalty rate they pay to Avanci. While $29 applies to all OEMs that joined in 2023, the rate ranges between $29 and $32 for the ones that joined after February 16, 2024. This calculation is performed in two ways: based on the number of days the licensee was part of the pool within the year and based on the quarters during which the licensee was active. Therefore, I assume that the vehicle sales are evenly distributed throughout the year.

4 Results

This section presents the key findings derived from the research method steps outlined above.

Beginning with the participation of the Avanci 5G licensors and licensees, Figure 1 illustrates the development of Avanci members from August 2023 until August 2024. The marks on the line indicate a new entrant. There has been an increase for both groups since the pool's inception on August 16, 2023. Initially, the pool started with 59 participants, of whom 58 were licensors and one was a licensee, namely Mercedes-Benz (Pohlmann, 2023). Over the next 12 months, the number of licensors grew by approximately 19% to 69, with OPPO being the last one that joined on July 31, 2024 (Ma, 2024). The licensees increased to 33 participants, with the Renault Group joining as last the OEM on August 26, 2024 (Avanci, 2024e). During the notable licensee increases in February, the VW Group joined the pool with 14 vehicle brands (Avanci, 2024g). Moreover, one third of the licensees joined before the royalty rate increased from $29 to $32 on February 16, 2024.

To compare the participation trends from another perspective, Table 2 provides the time-to-license distribution on a daily basis. While the 69 licensors joined within 350 days from the inception, it took the 33 licensees 376 days. The highest percentage of licensors (84.1%) joined immediately whereas most licensees (42.4%) joined after 188 days. Thus, the licensors' average time-to-license of 32.2 days is significantly lower than the 182.7 days for the licensees.

In addition to the previous results, the SEP coverage marks one of the research highlights. Figure 2 illustrates the

[17] Toyota's acquisition of 5G SEPs from OPPO is not taken into consideration.

[18] Saab produces defense vehicles. Iseki, Kobelco, MAN, Navistar, Scania, VW Truck & Bus and Yanmar manufacture industrial and agricultural vehicles but Statista (2023) identifies Deere & Company, CNH Industrial, Kubota, AGCO Corporation and Claas as global market leaders in this industry by revenue.

[19] Statista Passenger Cars Worldwide

[20] Scout will not start their production until 2026, and Cupra did not publish their revenue for 2023 or 2024. Therefore, I do not add them.

[21] I assume that the 66% share also applies for the manufacturers of agricultural and industrial vehicles.

[22] Iseki, Kobelco, Saab and Yanmar did not publish their unit sales for 2023.

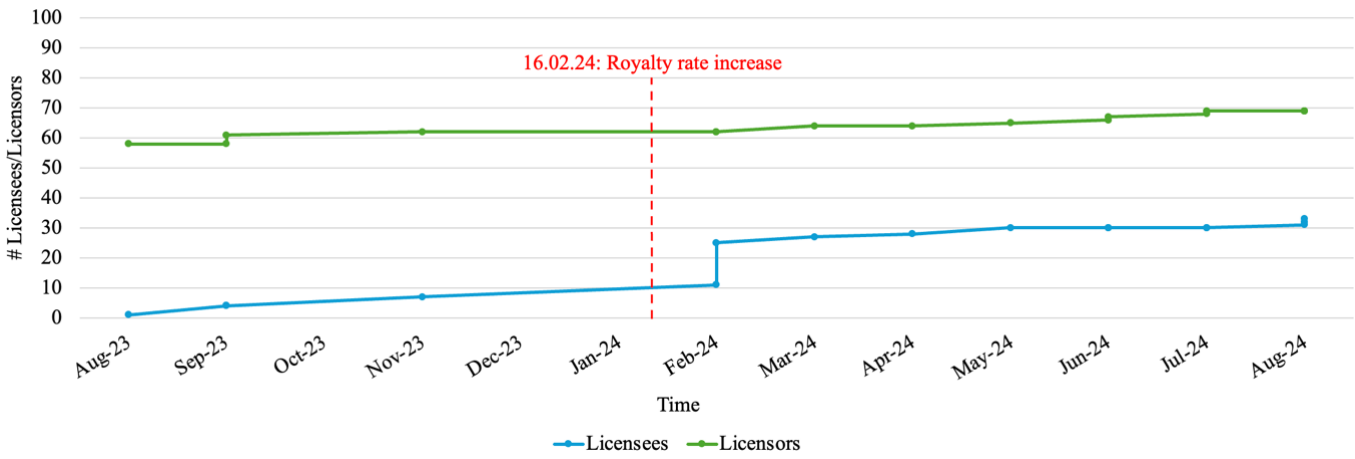

Figure 1: Number of Avanci 5G licensees and licensors (monthly)

Table 2: Time-to-license overview for licensees and licensors (daily)

Days	# Licensees	Share	# Licensors	Share
0	1	3.0%	**58**	**84.1%**
23	3	9.1%		
42			3	4.3%
83	3	9.1%	1	1.4%
181	4	12.1%		
188	**14**	**42.4%**		
222	2	6.1%	2	2.9%
256	1	3.0%		
276	2	6.1%	1	1.4%
300			1	1.4%
302			1	1.4%
344			1	1.4%
350			1	1.4%
356	1	3.0%		
372	1	3.0%		
376	1	3.0%		

development of the 2G-5G SEP coverage of the Avanci 5G pool from August 2023 to August 2024 at patent- and patent family level. The number of patents included in the Avanci 5G pool rose from 151,469 to 165,428 at patent level and from 58,160 to 64,559 at patent family level. Consequently, both coverages developed almost in parallel. While the coverage at patent family level increased from 77% to 86%, the coverage at patent level grew from 82% to 90%. That said, there was no development between November 2023 and June 2024. Two larger increases can be observed between October and November 2023 and June and July 2024. During the first jump, Intel joined on October 11, 2023, with 2,191 patents at patent level and 1,147 patents at patent family level (Lovatt, 2023). During the second one, OPPO signed an Avanci license with 11,583 patents at patent- and 5,204 patents at patent family level on July 29, 2024 (Ma, 2024).

The next assessment estimates the licensee coverage of the Avanci 5G pool. Globally, the combined connected car revenue of 129 major car manufacturers is projected to reach approximately USD 1,371 billion in 2024. 79 of these companies are headquartered in Asia, while 39 are in Europe and 11 in the USA. The 25 carmakers participating in Avanci are also headquartered on those three continents, but their distribution is significantly different. Avanci's licensees are concentrated in Europe (15) and the USA (6), with only 4 in Asia. Despite the generally large number of 56 OEMs in China and 11 in Japan, none hold an Avanci 5G license. In contrast, Europe, especially countries like Sweden und Germany, show high participation rates.

Besides the geographic coverage, Figure 3 presents the licensee revenue coverage development monthly. With the participation of multiple carmakers, the cumulative projected connected car revenue of the Avanci licensees increased by more than nine times from USD 67 billion in August 2023 to USD 619 billion in August 2024. As a result, the coverage of the total project connected car market revenues increased from 4.9% to 45.2% over this one-year period. Three major jumps occurred in September 2023, November 2023 and February 2024. While the BMW Group joined in the first one (Avanci, 2023a), the Hyundai Motor Group joined in the second (Avanci, 2023b) and the General Motor Company (Avanci, 2024d) and Volkswagen Group (Avanci, 2024g) in the third one. Nonetheless, among the ten largest players, Toyota, Honda, Nissan, BYD and Tesla still not entered the pool. Appendix 3 illustrates the licensee coverage calculation.

The increasing licensee coverages leads to the total rev-

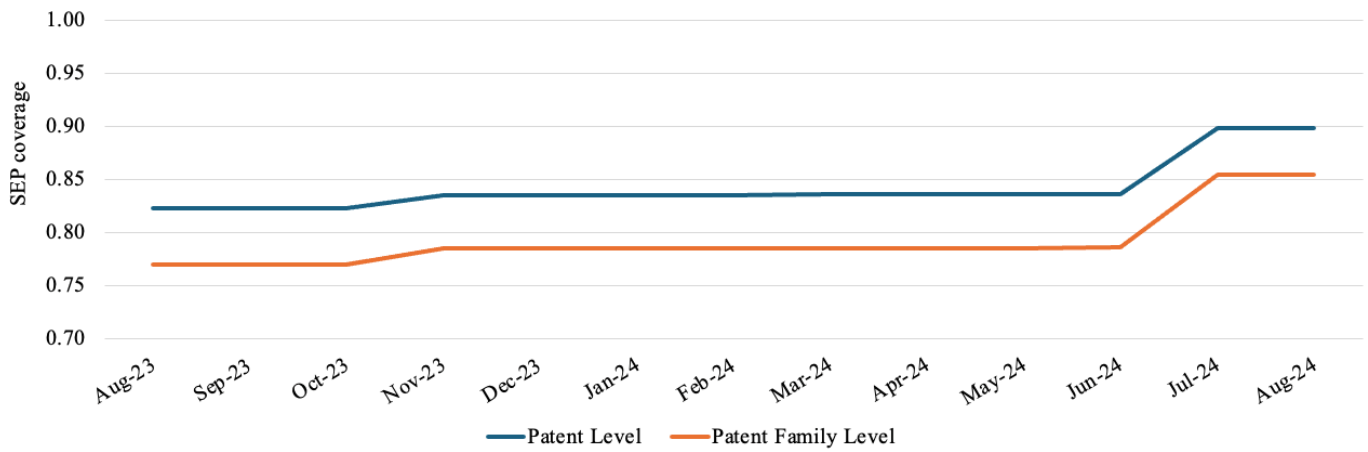

Figure 2: Avanci 2G-5G SEP coverage development (monthly)

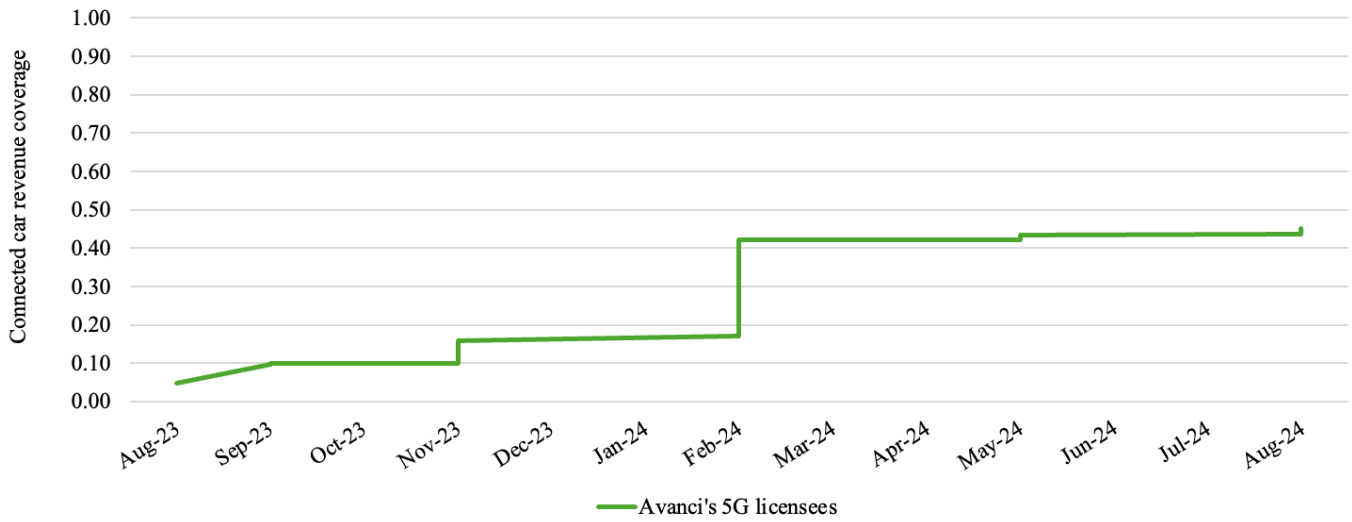

Figure 3: Avanci licensee coverage development (monthly)

enue generated by the Avanci 5G pool in 2023 and 2024. In 2023, the seven participating OEMs sold over 12.6 million connected vehicles and contributed around $54.5 million for the daily consideration and approximately $84.4 million for the quarterly consideration of their Avanci 5G license payments. With 25 new licensees in 2024, the vehicle sales and the pool administrator's revenue will increase significantly. Based on the estimated average growth rate of 2.3% in unit sales for all OEMs, Avanci's licensees are projected to sell over 29.7 million connected vehicles in 2024.[23] Brands like Volkswagen, Hyundai, Kia and Mercedes-Benz have the highest sales. As a result, depending on the rates applied to the licensees that joined after the price increase in February 2024, the revenue is expected to range between $495.5 million and $515.6 million for the daily consideration. This represents a year-over-year growth of 807% to 946%. In contrast, the quarterly calculation projects revenues of $539.1 million and $564.1 million for 2024, indicating an increase of 539% to 669%. Appendix 4 illustrates the Avanci revenue calculations for 2023 and 2024.

Besides the Avanci specific data and results, there are other observations around court data related to automotive connectivity and the broader SEP licensing ecosystem. First, it is important to note that there are no ongoing court cases where an Avanci licensor acts as a plaintiff again a potential licensee as a defendant.

Only one recent case, which was dismissed on July 15, 2024, involved Avanci and one of its members as defendants (Fancourt, 2023). In that case, Tesla tried to get a preferential 5G rate and raised concerns about the FRAND terms of the patent pool. While it was dismissed because the court did not find serious issues with Avanci or InterDigital, Tesla also failed to establish a sufficient legal basis.

[23] This number might be even higher because there are no vehicle sales data of Iseki, Kobelco, Saab and Yanmar.

In the other two ongoing cases, Intellectual Ventures is suing Volvo (Waldrop, 2023), an Avanci licensee, and Tesla (Siegmund, 2024) for various patent infringements at the Western Texas District Court. Appendix 5 provides an overview of the relevant court cases.

5 Discussion

Patent pools play a critical role in navigating complex intellectual property licensing ecosystems of technology-intensive industries. They provide a single point of negotiation for multiple licensors and licensees. This thesis provides a comprehensive analysis of the Avanci 5G pool, focusing on participation trends, SEP and licensee coverage, and revenue generation.

This section discusses the key findings, examines their theoretical and practical implications, and compares them to the existing literature. It also highlights the limitations of this study and offers recommendations for future research.

I collect the relevant data on the participating members and extract and prepare the ETSI database to retrieve all declared SEPs at patent- and patent family level. I calculate the SEP coverage with the share of SEPs included in the Avanci pool relative to the total number of SEPs declared to ETSI. Furthermore, I estimate the licensee coverage with a list of 129 global car manufacturers and use the licensees' unit sales to project the total revenues of the Avanci 5G pool. These steps lead to several relevant findings that can be interpreted and compared with the literature.

My first result highlights the extending participation of patent owners and OEMs in the Avanci 5G pool. The continuous growth of the pool from August 2023 to August 2024 underscores the rising importance of patent pools for SEP licensing within the IoT, particularly the technology-intensive automotive industry. This attractivity for both parties may result from the various benefits patent pools offer, ranging from lower transaction costs to reduced litigation risks. Moreover, this trend underscores a shift towards more collaborative innovation and shared resources with potential network effects in IPR management.

Furthermore, as the Avanci pool licenses multiple OEMs, it enables a more uniform and rapid implementation of the 5G technology within the industry. Avanci's strategy to engage in the IoT and 5G rollout matches the technology's remarkable development and impact highlighted by the Statista (2024b) report and Lanctot et al. (2023, pp. 4–16).

The overall participation developments align with the increasing significance of standardization, SEPs and patent pools, especially through the emergence of 5G, mentioned by the European Commission (2024b) and Lanctot et al. (2023, pp. 4–16). The need to handle the complexities of the licensing ecosystem, as illustrated by Pohlmann (2017), also support the interest in joining the Avanci pool.

Despite the general increase in participation, the analysis reveals a critical discrepancy in the time-to-license assessment among members. While most licensors joined the pool immediately at its inception, it took considerably longer for the licensees to join, with only one initial participant. This suggests different motivations and barriers for the members to join. On the one hand, licensors may prefer to join quickly to monetize their patents, cover R&D costs and promote broad technology adoption, aligning with the objectives mentioned by Dewatripont and Legros (2013, p. 928). Licensees, on the other hand, might take a more cautious approach due to cost considerations. They may weigh the benefits of joining the pool against negotiating individual bilateral licensing agreements with the patent owners individually. Market readiness and other aspects of the company's strategic planning, such as product launch cycles, may also be reasons for later participation.

The impact of major licensor joiners like OPPO or Intel, that hold thousands of 2G-5G SEPs, is questionable. While their participation enhances the pool's portfolio and patent coverage, it also strengthens the bargaining power and dominant position of Avanci. Smaller and later joiners could experience less favorable terms and decreased bargaining power. In fact, as already done for the 4G pool in 2022, Avanci may argue with a higher value proposition for vehicle manufacturers and increase the pool's royalty rate even further. This might result in higher vehicle production costs and end-product prices. Regarding the current royalty rate, it is important to note that most licensees joined after the royalty rate increase on February 16, 2024. However, the *"other conditions"* (Avanci, 2024f) applicable to later entrants, depend on the circumstances and indicate that not all licensees who joined after the increase have to pay the higher rate.

The late adoption by OEMs may be consistent with the continued resistance to the Avanci 5G pool. Six major automotive groups issued a letter to express the concerns about the lack of transparency and FRAND conditions (ACEA, CLEPA, CEMA, CECE, EGMF, & FEM, 2024). A second letter from the Chinese State Administration for Market Regulation (2024) indicates potential disadvantages against the dominant pool and calls for monopoly risk assessments related to its licensing practices. In addition to the letters, the formation of alliances like the automotive LNG (Bundeskartellamt, 2024) reflects the manufacturers' perceived urge to counter balance Avanci's negotiation power.

Additionally, it is evident that Avanci only licenses to end-product manufacturers and not to component suppliers. This allows them to maintain high royalty rates due to the higher profit margins of OEMs. The literature outlines several risks regarding the competitiveness and FRAND compliance of this practice (European Commission 2017; Nazzini 2017, p. 21). Disputes over whether manufacturers or suppliers should license a technology, as seen in the Daimler vs. Nokia case (Raturi, 2021), might harm licensee participation. Both parties want the other to bear the licensing costs, but nothing has yet been published about the potential of sharing them.

My second result addresses the first part of the research question and demonstrates the increasing SEP coverage of the Avanci 5G pool at patent- and patent family level. Since

the pool's inception, the coverage increased to 86% at patent level and 90% at patent family level. This extensive coverage strengthens the role of Avanci as a "one-stop shop" for 2G-5G technology licensing within the connected vehicle industry. The pool enhances efficient licensing by reducing the need for multiple bilateral negotiations and providing a broad patent portfolio that aims to foster technology implementation. These benefits are attractive for licensors and licensees and might make Avanci "inevitable".

However, this raises a critical question: Will Avanci's coverage lead to a "lock-in" effect, making OEMs and patent owners dependent and hindering bilateral agreements and other licensing models to emerge? With its expanding coverage, the pool administrator might have the negotiation power to justify higher licensing rates and achieve more favorable terms and conditions. Furthermore, other licensing models might struggle to compete as they will not be able to offer a similar coverage. Avanci has the potential to become the default licensing option for OEMs and patent owners, which reduces their motivation to explore alternatives or negotiate individually. This aligns with the anti-competitive concerns raised by the SEPs Expert Group (2021, p. 27) and Chinese State Administration for Market Regulation (2024). If Avanci maintains its approach of licensing only end-product manufacturers, it will also force OEMs and suppliers to agree on shared licensing costs to resolve the previously mentioned financial disputes in the long term. To avoid legal challenges around FRAND access to the technologies, the pool should ensure that its practices do not restrict competition or innovation within the automotive licensing ecosystem.

The questions brought up by ACEA, CLEPA, CEMA, CECE, EGMF, & FEM (2024, pp. 2–3) provide a good starting point for further discussion. They question the essentiality checks of the patents included and seek a justification for the 60% royalty rate increase compared to the 4G pool. Although Avanci (2024f) addresses some issues in their Q&A, the answers are general and lack transparent and detailed information on these aspects. Essentiality checks are particularly important as the European Commission (2023a, p. 25) stresses that only 20-30% of declared SEPs are actually essential, with only 10-20% for 5G technology.

In the interview with Jiang (2024), Avanci President Laurie Fitzgerald explained that the pool's conditions were established through input from multiple stakeholders on *"what set of licensing terms they could agree to"*. She also mentioned the platform costs and *"the value that cellular connectivity is going to bring to a 5G connected vehicle"* as contribution factors for the royalty rate. This approach does not directly align with the comparable license or top-down approach outlined by Bekkers et al. (2022, p. 5). Instead, it rather focuses on the perceived value and stakeholder consensus.

Given that many members of the automotive groups who signed the letter of complaint already hold an Avanci 5G license and may have provided input during the pool's creation, it is questionable where these concerns come from. That said, literature also states that SEP licensing frequently leads to disputes because FRAND commitments are *"inherently vague"* Geradin (2020, p. 29).

The letter of complaint could significantly impact the role and perception of lobbying in SEP licensing. Internal disputes might become more visible to the public and bring up questions on how parties behave in private negotiations compared to public statements. However, the letter can also be seen as a mechanism that ensures order and transparency within the industry. This might establish a precedent for groups in other industries to engage in more collaborative lobbying. The recently approved automotive LNG (Bundeskartellamt, 2024) will be an important aspect for that as well. OEMs have realized that they need to collaborate and combine their strengths to compete against influential patent pools on the licensor side.

For the second finding, it is important to note that there is an unproven essentiality verification of the SEPs declared to ETSI. This could lead to an overestimation of the patent coverage due to "over-declaration" by patent holders.

My third result presents the growing licensee coverage of the Avanci 5G pool and addresses the second part of the research question. The finding shows the notably different geographic distribution of the Avanci licensees compared to the 129 global car manufacturers. The low participation of only four Asian OEMs and the absence of Chinese and Japanese ones is striking. In contrast, Europe and the USA have much higher participation rates, with around 40% and 55%, respectively.

The skewed geographic distribution may reflect regional differences in IPR strategies and patent pool participation. The lack of interest of major Chinese and Japanese vehicle manufacturers suggests broader geopolitical and economic influences. For Avanci, the absence of numerous major OEMs creates a market gap. Understanding regional participation is crucial to reach as many licensees as possible and fully capitalize on the pool's growth.

Avanci might struggle to attract Chinese licensees because they follow another licensing approach, as the absence of them is also observable in the Avanci 4G pool (Avanci, 2024a). Given that Counterpoint (2024a) expects China to lead the connected car market, followed by the USA, Japan and Germany, it is particularly important for the pool administrator to tackle this issue. This could include addressing specific concerns that Chinese OEMs have, ranging from dependencies on western-controlled patent pools to anti-competitive practices expressed by the letter of the Chinese State Administration for Market Regulation (2024). The letter seeks a fair licensing market and urges the pool to monitor monopoly risks. Chun (2024) illustrates that failure to do so could lead to potential penalties or a continued lack of market share in this critical market. While one might think that Avanci's focus and practices are regionally biased, it is important to note that their licensors are more equally distributed across the three continents, with the majority in Asia.

In addition to the geographic coverage, the estimated revenue coverage of the Avanci licensees shows substantial growth since the pool's inception. The coverage is expected

to rise from 4.9% to 45.2% of the global car market revenues, driven by an extensive increase in the cumulative projected revenues of Avanci's licensees for 2024. While major OEMs such as BMW, Hyundai, General Motors, and Volkswagen caused notable coverage jumps, global players like Toyota, Honda, Nissan, BYD, and Tesla have not yet joined the pool.

The rapid increment over the one-year period suggests that the Avanci 5G pool is becoming central to the automotive industry. This would align with the various financial and legal benefits a patent pool offers, as elaborated on by Dequiedt and Versaevel (2013, pp. 59–69). For the OEMs, participating in efficient licensing agreements is a key driver of technological success. Nevertheless, major OEMs who decided not to join may seek more favorable terms to their specific needs and want to maintain control over their licensing partners and costs. Joining the pool could mean to agree on the same conditions as their competitors. However, the pool's growing SEP coverage and the resulting efficiency benefits are likely to increase the incentives to join.

While Precedence Research (2024) estimates that 35% of the connected car market are located in Asia, only 13% of the projected revenues of the current Avanci members will be in this region. This underscores the significant underrepresentation of Asian OEMs in the pool. If Avanci fails to cover this market, it could miss out on future growth.

My fourth result further reveals the increasing relevance and attraction of the Avanci 5G patent pool. Its extensive revenue jump is accelerated by the rapid expansion in the number of participating licensees since its inception. It is evident that Avanci is generally successful in attracting a broad range of OEMs.

Successful pools with a broad coverage will become unavoidable as licensing negotiations become more complex (Nikolic & Galli, 2021). However, Avanci's extensive revenue growth emphasizes that regulators need to ensure competition, aligning with the careful management to maintain a balanced ecosystem stressed by European Commission (2023b, pp. 1–3).

That said, I am not able to identify the individual rates of the licensees who joined the pool after the price increase and therefore only use two scenarios for the pool's revenue calculation. Assuming a 66% share of connected car revenues aligns with the forecast by Precedence Research (2024). However, this estimate could be biased, and the actual share might be even higher for Avanci's OEMs who demonstrate an above-average emphasis on connectivity through their pool participation.

My last result highlights the absence of legal cases with Avanci licensors acting as plaintiffs. This implies effective licensing terms and litigation risk reduction. As a result, the licensing environment of the Avanci 5G pool appears more secure, stable and predictable not only for potential entrants but also regulators and legal bodies. This security further enhances Avanci's position and reputation within the industry. However, the pool must be aware of future legal risks and FRAND compliance. Outcomes of key cases can impact negotiations and litigation strategies.

The recently dismissed case between Tesla and InterDigital and Avanci is of special interest. The court's decision to reject Tesla's concerns on Avanci's FRAND terms demonstrates the position of the pool (Müller, 2024c). It appears that courts may be hesitant to interfere with Avanci's established licensing model. Additionally, the dismissal indicates the limited ability of OEMs to negotiate alternative licensing conditions. However, the issue of whether courts can impose FRAND terms for patent pools remains unresolved. This outcome might prevent individual licensees from challenging the pool but may encourage them to create other counter pools to compete.

In addition to my results, there are other aspects that need to be discussed in the context of the impact of regulatory developments and new strategic positions in the automotive SEP licensing landscape.

The proposed EU SEP regulations published by the European Commission (2023b) and the UK SEPs Resource Hub promoted by the UK Government (2024) highlight two distinctly different approaches. The EU's interventionist strategy seeks a regulatory framework to address transparency and FRAND concerns. This indicates an effort to prevent market abuse and ensure fair SEP licensing. However, EU's framework could lead to stricter oversights of SEP licensing, which extends negotiation processes and reduces flexibility. Companies would therefore need to invest in legal and compliance resources to navigate the new regulations.

In contrast, the UK aims to provide a centralized resource hub which assists UK businesses in navigating the complex SEP licensing ecosystem. The approach promotes self-regulation and market-driven solutions. It will also make it easier for unfamiliar companies to access and understand information related to SEPs. Thereby, the UK lowers the entry barriers and encourage a market, where informed players can negotiate more effectively. That could lead to a dominant position of more established companies. However, the UK resource hub's effectiveness will be limited as companies might still face challenges when dealing with different jurisdictions.

It is questionable which approach better balances the interests of licensors and licensees while fostering technological innovation and ensuring fair competition. Depending on the performance of the respective markets, these approaches might set a precedent for other regions. However, different regulations create a fragmented environment where companies like Avanci must handle different rules and expectations depending on the region.

Although Avanci is headquartered in the USA, it's members also operate in the EU and UK. Within the EU, Avanci might need to increase operational oversight and slow down some processes to ensure that its practices align with the new regulations. These new regulations could for example force them to also license component suppliers. As the UK approach allows more flexibility, Avanci might favor it. The pool's approach to incorporate feedback from various stakeholders (Jiang, 2024) and creating a *"market-driven pricing"* model (Avanci, 2024f) can already be linked to the UK's ap-

proach.

The last important topic are the shifts in how OEMs are approaching IPRs and licensing strategies. We can see this in Toyota's recent acquisition of SEPs from OPPO (Müller, 2024d) and the formation of the automotive LNG (Bundeskartellamt, 2024).

On the one hand, the move by Toyota might be the response to increasing complexities and costs of SEP licensing. With the acquisition, the vehicle manufacturer enhances its portfolio and bargaining power. This step towards autonomy allows the use of the acquired patents in cross-licensing deals to reduce dependencies and costs. Regarding patent pool participation, this means that the OEM could enter as a licensor and licensee at the same time.[24] These advantages might encourage other OEMs to approach a similar strategy.

On the other hand, the automotive LNG tries to push back against the risk of high licensing costs and unfavorable terms of dominant pools like Avanci. As vehicles become more technology-intensive, this can significantly impact the profitability of OEMs. Traditionally, Avanci seemed to be driving the SEP licensing terms within the automotive industry. The counter pool might introduce a new dynamic that could rebalance the automotive SEP ecosystem. However, the success of the LNG will depend on its ability to attract members and navigate regulatory, financial and transparency challenges.

Another important part of this section are the limitations of this research. In general, the focus only on the Avanci 5G patent pool illustrates a methodological constraint. As other industries and markets may have different patent pools and licensing structures, my results should be transferred to other contexts with caution. The findings are applicable to Avanci and the automotive industry, but other sectors might not experience the same outcomes. Furthermore, it is important to keep in mind that this study analyzed the pool until August 31, 2024. The automotive industry has already evolved rapidly during the one-year period since the pool's inception, which indicates a high probability of significant further developments that need to be evaluated when reviewing the outcomes of this thesis.

As a result, I have some suggestions for future research. If the scope and available resources allow, I recommend improving the completeness and accuracy of the dataset to address some of the limitations of my research. Regarding the Avanci pool, it would be interesting to understand, why major OEMs refuse to join. Also, the industry's perception of the impact of new trends, including the remaining resistance, the emergence of the automotive LNG and proposed regulatory frameworks, could be valuable. However, this should also be compared with the dynamics of patent pools in other industries.

[24] Hyundai and Kia recently joined the Versatile Video Coding patent pool as licensors, becoming the first OEMs to participate as licensors in a major patent pool (Morris, 2024).

6 Conclusion

The emergence of new technologies that heavily rely on standardization significantly increases the complexity of SEP licensing, highlights the importance of patent pools and underscores the need to evaluate their participation dynamics and operational achievements.

This thesis analyzes the participation trends in the Avanci 5G vehicle pool, focusing on the development of the pool's patent and licensee coverages since its inception on August 16, 2023. The study uses data on the pool members and relevant SEPs declared to ETSI to estimate the pool's coverages and total revenue.

Increased participation by patent holders and OEMs has led to a substantial growth in both coverages over the past year. This illustrates the extensive significance of the pool in the technology-intensive automotive industry, resulting in a notable growth in estimated revenue for 2024. However, the pool's inability to attract major Asian OEMs, particularly from China, suggests their favor for other licensing strategies. Also, perceived anti-competitive risks, such as "lock-in" effects and the restriction of bilateral agreements, could be caused by the dominance of the pool.

Therefore, future research should include in-depth case studies of large companies that have not joined the pool to understand alternative licensing strategies and regional differences. Furthermore, the impact of geopolitical dynamics on patent pool participation and the SEP licensing landscape needs to be explored. Expert opinion gathered through industry and academic interviews could further support this. Given this study's focus on the automotive industry, comparative research across different sectors would help to understand the effectiveness, challenges, and impact of patent pools more broadly.

Looking ahead, emerging technologies, evolving regulatory frameworks, and new licensing strategies will continue to reshape the licensing ecosystem. Stakeholders must navigate these challenges collaboratively. Pool administrators need transparent and flexible licensing terms to attract a wider range of participants and adapt to new market conditions. The long-term success of a pool will depend on how well it can balance the interests of licensors and licensees without abusing its power. Regulators must promote an approach that encourages competition and innovation without creating overwhelming regulatory burden. An adaptive regulatory framework that evolves with technological advancements could be the solution. Therefore, the formation of working groups that include all stakeholders could benefit the discussion and negotiation of specific terms.

References

ACEA, CLEPA, CEMA, CECE, EGMF, & FEM. (2024). Multi-Association Letter to Avanci. https://www.acea.auto/files/Multi-Assication_Letter_to_Avanci.pdf

Avanci. (2023a). BMW Group joins Avanci 5G Connected Vehicle Licensing Program. https://www.avanci.com/2023/09/18/bmw-group-joins-avanci-5g-connected-vehicle-licensing-program/

Avanci. (2023b). Momentum continues for Avanci and its independent, global licensing solutions as Hyundai and Kia sign 5G vehicle agreements. https://www.avanci.com/2023/11/07/momentum-continues-for-avanci-and-its-independent-global-licensing-solutions-as-hyundai-and-kia-sign-5g-vehicle-agreements/

Avanci. (2024a). Avanci 4G Vehicle. https://www.avanci.com/vehicle/4gvehicle/

Avanci. (2024b). Avanci 5G Vehicle. https://web.archive.org/web/20240826071131/https://www.avanci.com/vehicle/5gvehicle/

Avanci. (2024c). Avanci 5G Vehicle. https://www.avanci.com/vehicle/5gvehicle/

Avanci. (2024d). Avanci Announces 5G Vehicle License Agreement with General Motors. https://www.avanci.com/2024/02/13/avanci-announces-5g-vehicle-license-agreement-with-general-motors/

Avanci. (2024e). Avanci Launches 5G Connected Vehicle Licensing Program. https://www.avanci.com/2023/08/16/avanci-launches-5g-connected-vehicle-licensing-program/

Avanci. (2024f). Avanci Vehicle. https://www.avanci.com/vehicle/

Avanci. (2024g). Avanci Welcomes the Volkswagen Group to its 5G Vehicle Licensing Program. https://www.avanci.com/2024/02/20/avanci-welcomes-the-volkswagen-group-to-its-5g-vehicle-licensing-program/

Avanci. (2024h). What Is Avanci? https://www.avanci.com/

Barnett, J. M. (2018). Has the Academy Led Patent Law Astray? *Berkeley Technology Law Journal*, *32*(5), 1321–1324. https://doi.org/10.15779/Z388P5V91H

Baron, J., & Ménière, Y. (2014). Patent Pools. *Selected Quantitative Studies of Patents in Standards*, 55. https://doi.org/10.2139/ssrn.2457064

Baron, J., & Pohlmann, T. (2018). Mapping standards to patents using declarations of standard-essential patents. *Journal of Economics & Management Strategy*, *27*(3), 504–534. https://doi.org/10.1111/jems.12255

Baron, J., & Schmidt, J. (2019). Technological Standardization, Endogenous Productivity and Transitory Dynamics. *Banque de France*, *503*, 5. https://doi.org/10.2139/ssrn.2478707

Bekkers, R., Raiteri, E., Martinelli, A., Tur, E. M., & Thumm, N. (2020). Landscape Study of Potentially Essential Patents Disclosed to ETSI. Pilot Study for Essentiality Assessment of Standard Essential Patents. https://data.europa.eu/doi/10.2760/313626

Bekkers, R., Tur, E. M., Henkel, J., van der Vorst, T., Driesse, M., & Contreras, J. L. (2022). Overcoming Inefficiencies in Patent Licensing: A Method to Assess Patent Essentiality for Technical Standards. *Research Policy*, *51*(10), 5. https://doi.org/10.1016/j.respol.2022.104590

Bellflamme, P. (2002). Coordination on Formal vs. De Facto Standards: A Dynamic Approach. *European Journal of Political Economy*, *18*, 153–176. https://www.sciencedirect.com/science/article/pii/S0176268001000738?via%3Dihub

Blind, K., & Thumm, N. (2004). Interrelation Between Patenting and Standardisation Strategies: Empirical Evidence and Policy Implications. *Research Policy*, *33*(10), 1583–1598. https://doi.org/10.1016/j.respol.2004.08.007

Brazier, T. (2024). A Closer Look at the New UKIPO SEP Resource Hub. https://www.iam-media.com/hub/sepfrand-hub/2023/article/closer-look-the-new-ukipo-sep-resource-hub

Bundeskartellamt. (2024). Proposed creation of a framework for negotiating licence agreements for standard essential patents (SEPs) through an "Automotive Licensing Negotiation Group" (ALNG). https://www.bundeskartellamt.de/SharedDocs/Publikation/EN/Others/B4_136_23_Vorsitzendenschreiben.html?nn=52004

Chiesa, V., Manzini, R., & Toletti, G. (2002). Standard-setting processes: evidence from two case studies. *R&D Management*, *32*, 431–450. https://onlinelibrary.wiley.com/doi/10.1111/1467-9310.00274

Chun, Z. (2024). A Deep Dive into China's Three-Letters-One-Notice System as Compliance Challenges Emerge for Patent Pools. https://www.iam-media.com/hub/sepfrand-hub/2023/article/deep-dive-chinas-three-letters-one-notice-system-compliance-challenges-emerge-patent-pools

Cohen, L., Gurun, U., & Kominers, S. (2018). Patent Trolls: Evidence from Targeted Firms. http://dx.doi.org/10.2139/ssrn.2464303

Counterpoint. (2024a). Global Connected Car Sales to Exceed 500 Million in 2024-2030. https://www.counterpointresearch.com/insights/global-connected-car-sales-to-exceed-500-million-in-2024-2030/

Counterpoint. (2024b). Two-thirds of Cars Sold in Q3 2023 Featured Embedded Connectivity. https://www.counterpointresearch.com/insights/global-connected-car-sales-q3-2023/

David, P., & Greenstein, S. (1990). The Economics of Compatibility Standards: An Introduction to Recent Research. *Economics of Innovation and New Technology*, 3–41. https://doi.org/10.1080/10438599000000002

Dequiedt, V., & Versaevel, B. (2013). Patent Pools and dynamic R&D incentives. *International Review of Law and Economics*, *36*, 59–69. https://doi.org/10.1016/j.irle.2013.04.009

Dewatripont, M., & Legros, P. (2013). "Essential" patents, FRAND royalties and technological standards. *The Journal of Industrial Economics*, *61*(4), 928. https://doi.org/10.1111/joie.12033

Dini, G. (2024). Aggregator-Arranged Video Patent Deal Brings Licensing Efficiency. https://www.sisvel.com/news/aggregator-arranged-video-patent-deal-brings-licensing-efficiency/

ETSI. (2024a). Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards. https://www.etsi.org/standards-search?search=SR+000+314#page=1&search=SR%20000%20314&title=1&etsiNumber=1&content=0&version=0&onApproval=1&published=1&withdrawn=1&historical=0&isCurrent=1&superseded=1&startDate=1988-01-15&endDate=2024-07-31&harmonized=0&keyword=&TB=&stdType=&frequency=&mandate=&collection=&sort=1

ETSI. (2024b). Intellectual Property Rights Policy and IPR Online Database. https://www.etsi.org/intellectual-property-rights

ETSI. (2024c). Why Standards. https://www.etsi.org/standards/why-standards?jjj=1721032782036

European Commission. (2014). Patents and Standards – A Modern Framework for IPR-Based Standardization - Final Report, 67–68. https://data.europa.eu/doi/10.2769/90861

European Commission. (2017). Setting out the EU approach to Standard Essential Patents. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC0712&from=EN

European Commission. (2023a). Economic impact assessment on Standard Essential Patents (SEPs) - Support study. https://data.europa.eu/doi/10.2873/19262

European Commission. (2023b). Regulation of the European Parliament and of the Council on Standard Essential Patents and Amending Regulation. https://single-market-economy.ec.europa.eu/publications/com2023232-proposal-regulation-standard-essential-patents_en

European Commission. (2024a). Intellectual Property – New Framework for Standard-Essential Patents - Feedback and Statistics. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13109-Intellectual-property-new-framework-for-standard-essential-patents/feedback_en?p_id=28414115

European Commission. (2024b). Standard Essential Patents. https://single-market-economy.ec.europa.eu/industry/strategy/intellectual-property/patent-protection-eu/standard-essential-patents_en

Fancourt, J. (2023). Tesla vs. InterDigital and Avanci. https://www.bailii.org/ew/cases/EWHC/Ch/2024/1815.pdf

Geradin, D. (2020). SEP Licensing after Two Decades of Legal Wrangling: Some Issues Solved, Many Still to Address. *Antitrust Chronicle*, *3*, 25–38. https://www.competitionpolicyinternational.com/category/winter-2020-volume-3-number-2/

Geradin, D., & Katsifis, D. (2021). End-product- vs Component-level Licensing of Standard Essential Patents in the Internet of Things Context. https://doi.org/10.2139/ssrn.3848532

Houldsworth, A. (2024). Sisvel and RPX strike major video coding patent deal. https://www.iam-media.com/article/sisvel-and-rpx-strike-major-video-coding-patent-deal

Jiang, J. (2024). Dialogue with Avanci Vehicle President Laurie Fitzgerald: Avanci, an Independent Platform Provides a Licensing Option. http://www.chinaiptoday.com/post.html?id=2103

Kanevskaia, O. (2020). The Need for Multi-disciplinary Education About Standardization. *Sustainable Development*, 161–178. https://doi.org/10.1007/978-3-030-28715-3_10

Khanh, Q. V., Hoai, N. V., Manh, L. D., Le, A. N., & Jeon, G. (2022). Wireless Communication Technologies for IoT in 5G: Vision, Applications, and Challenges. *Wireless Communications and Mobile Computing*, 1–3. https://doi.org/10.1155/2022/3229294

Klos, M. (2020). Federal Cartel Office Issues Opinion in Connected Cars Case. https://www.juve-patent.com/cases/federal-cartel-office-issues-opinion-in-connected-cars-case/

Koh, L. (2019). Continental Automotive Systems v. Avanci. https://casetext.com/case/contl-auto-sys-v-avanci-llc-4

Lanctot, R., Kendall, P., & Koehler, A. (2023). 5G Connected Cars: A Transformative Value Proposition. *Tech Insights*, 4–16. https://www.avanci.com/5g-report/

Lavenue, L. M., Myles, J. M., & Evans, R. (2023). IP Transactions, Professional Perspective - Future of SEP Licensing in the Auto Industry. *Bloomberg Law*. https://www.finnegan.com/en/insights/articles/ip-transactions-professional-perspective-future-of-sep-licensing-in-the-auto-industry.html

Lemley, M., & Shapiro, C. (2013). A Simple Approach to Setting Reasonable Royalties for Standard-Essential Patents. *Berkeley Technology Law Journal*, *28*, 1137. https://doi.org/10.2139/ssrn.2243026

Lerner, J., & Tirole, J. (2015). Standard-Essential Patents. *Journal of Political Economy*, *123*(3), 547–586. https://doi.org/10.1086/680995

Lovatt, S. (2023). Intel Signs Up as Licensor to Avanci 4G and 5G Connected Car Programmes. https://www.iam-media.com/article/intel-signs-licensor-avanci-4g-and-5g-connected-car-programmes

Ma, M. (2024). OPPO joins Avanci 5G. https://michael7924.substack.com/p/oppo-joins-avanci-5g?r=12l5b8&utm_campaign=post&utm_medium=web&triedRedirect=true

Martinez, J. (2019). FRAND as Access to All versus License to All. *Journal of Intellectual Property Law & Practice*, *14*(8), 649. https://doi.org/10.1093/jiplp/jpz075

Morris, A. (2024). Hyundai Makes Move in Goal to Be Future SEP Licensor. https://www.iam-media.com/article/hyundai-makes-move-in-goal-be-future-sep-licensor

Mossoff, A. (2023). Via Licensing, MPEG-LA Merger Shows Patent Pools Have Come of Age. https://www.iam-media.com/article/licensing-mpeg-la-merger-shows-patent-pools-have-come-of-age

Muller, F. (2022). Frustrated over Avanci patent pool's market penetration, critics ask DOJ to revisit Business Review Letter–but their disagreement with German patent law is no U.S. antitrust problem. *Foss Patents*. http://www.fosspatents.com/2022/10/frustrated-over-avanci-patent-pools.html

Muller, F. (2023). New Mercedes-Benz statement: despite Avanci 5G license, EU should enact regulation on standard-essential patents in accordance with DG GROW's legislative proposal. *Foss Patents*. http://www.fosspatents.com/2023/08/new-mercedes-benz-statement-despite.html

Müller, F. (2024a). About ip fray. https://ipfray.com/about/

Müller, F. (2024b). Philips gets 5G SEPs from OPPO, probably under recent settlement agreement and following OPPO's recent Toyota deal. https://ipfray.com/philips-gets-5g-seps-from-oppo-probably-under-recent-settlement-agreement-and-following-oppos-recent-toyota-deal/

Müller, F. (2024c). Tesla tells UK court it has Avanci 4G license but wants preferential 5G rate, estimates Avanci's coverage at 80%. https://ipfray.com/tesla-tells-uk-court-it-has-avanci-4g-license-but-wants-preferential-5g-rate-estimates-avancis-coverage-at-80/

Müller, F. (2024d). Toyota acquired dozens of 5G SEPs from OPPO to equip itself for cross-licensing and countersuits. https://ipfray.com/toyota-acquired-dozens-of-5g-seps-from-oppo-to-equip-itself-for-cross-licensing-and-countersuits/

Nazzini, R. (2017). Level Discrimination and FRAND Commitments under EU Competition Law. *World Competition*, *40*(2), 21. https://doi.org/10.54648/woco2017015

Nikolic, I., & Galli, N. (2021). A Look Into the IoT Future of SEP Licensing. *CPI IP Columns*, 1–8. https://ssrn.com/abstract=3848041

Nikolic, I., & Galli, N. (2022). Patent Pools in 5G: The principles for facilitating pool licensing. *Telecommunications Policy*, *46*(4). https://doi.org/10.1016/j.telpol.2021.102287

OECD. (2019). Licensing of IP Rights and Competition Law. *OECD Roundtables on Competition Policy Papers*, *230*, 24–36. https://doi.org/10.1787/6a74221e-en

Pohlmann, T. (2015). Landscaping Standard-Essential Patents. https://www.iam-media.com/article/landscaping-standard-essential-patents

Pohlmann, T. (2017). Patents and Standards in the Auto Industry. https://www.iam-media.com/article/patents-and-standards-in-the-auto-industry

Pohlmann, T. (2023). Avanci's new 5G Vehicle Program – A One Stop Shop? https://www.linkedin.com/pulse/avancis-new-5g-vehicle-program-one-stop-shop-tim-pohlmann/

Precedence Research. (2024). Connected Cars Market Size, Share, and Trends 2024 to 2034. https://www.precedenceresearch.com/connected-cars-market

Rabbi, N. F. (2020). Introduction to 5G. *Islamic University of Technology*, *1*. http://dx.doi.org/10.13140/RG.2.2.14027.00804

Raturi, P. (2021). Limiting SEP Licensing to End Product Manufacturers to the Exclusion of Component Suppliers- Abuse of Dominance? *Oxford Business Law Blog*. https://blogs.law.ox.ac.uk/business-law-blog/blog/2021/08/limiting-sep-licensing-end-product-manufacturers-exclusion-component

Schüll, G. (2024). Bundeskartellamt Allows Cooperation Between Automotive Companies in SEP Licensing. *Cohausz & Florack*. https://www.cohausz-florack.de/en/blog/article/bundeskartellamt-allows-cooperation-between-automotive-companies-in-sep-licensing/

Schwartz, D., & Kesan, J. (2014). Analyzing the Role of Non-Practicing Entities in the Patent System. *Cornell Law Review*, *99*(2), 426–456. https://doi.org/10.2139/ssrn.2117421

SEPs Expert Group. (2021). *Contribution to the Debate on SEPs* (tech. rep. No. 27). European Commission. https://ec.europa.eu/docsroom/documents/45217

Shan, G. (2024). Toyota acquires dozens of communications SEPs from OPPO to address patent dilemma of automakers. *Intellectual Property Finance*. https://mp.weixin.qq.com/s/r_zB9beG7y4i2GWBxXHOhg

Siegmund, M. (2024). Intellectual Ventures vs. Tesla. *Western Texas District Court*. https://s3.documentcloud.org/documents/24547997/24-04-12-iv-v-tesla-complaint.pdf

State Administration for Market Regulation. (2024). The State Administration for Market Regulation Has Legally Reminded and Urged Avanci to Address the Monopoly Risks of Its Patent Pool. https://www.samr.gov.cn/xw/zj/art/2024/art_ec1c9ce3d71a4a5baf853d430a3b5667.html

Statista. (2020). Share of New Vehicles Shipped Worldwide with Built-In Connectivity in 2020 and 2030. https://www.statista.com/statistics/1276018/share-of-connected-cars-in-total-new-car-sales-worldwide/

Statista. (2022). Connected Vehicle Fleet from 2025 to 2035, by Key Region. https://www.statista.com/statistics/442529/connected-car-market-size-by-region/

Statista. (2023). Selected Farm Machinery Manufacturers Worldwide in FY 2022, Based on Revenue. https://www.statista.com/statistics/461428/revenue-of-major-farm-machinery-manufacturers-worldwide/

Statista. (2024a). Number of Cars Sold Worldwide from 2010 to 2023, with a 2024 Forecast. https://www.statista.com/statistics/200002/international-car-sales-since-1990/

Statista. (2024b). Number of Internet of Things (IoT) Connections Worldwide from 2022 to 2023, with Forecasts from 2024 to 2033. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

Sterzi, V., Rameshkoumar, J.-P., & Van Der Pol, J. (2020). *Non-practicing entities and transparency in patent ownership in Europe* (tech. rep. No. 4). Bordeaux Economic Working Paper. https://ideas.repec.org/p/grt/bdxewp/2020-10.html

UK Government. (2024). One-Stop SEPs Resource Hub Launched by UK IPO. https://www.gov.uk/government/news/one-stop-seps-resource-hub-launched-by-uk-ipo

Via Licensing Alliance. (2023). Via Licensing and MPEG LA Unite to Form Via Licensing Alliance, the Largest Patent Pool Administrator in the Consumer Electronics Industry. https://www.via-la.com/via-licensing-and-mpeg-la/

Via Licensing Alliance. (2024). About Via LA. https://www.via-la.com/about-2/

Waldrop, J. (2023). Intellectual Ventures vs. Volvo. *Western Texas District Court*. https://s3.documentcloud.org/documents/23837560/23-06-07-intellectual-ventures-v-volvo-patent-infringement-complaint.pdf

List of Abbreviations

ETSI European Telecommunications Standards Institute

EU European Union

FRAND Fair, Reasonable and Non-Discriminatory

IoT Internet of Things

IPR Intellectual Property Right

LNG Licensing Negotiation Group

NPE Non-Practicing Entity

OEM Original Equipment Manufacturer

R&D Research & Development

SEP Standard-Essential Patent

SSO Standard-Setting Organization

UK United Kingdom

USA United States of America

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Generative AI-Enabled Music Generation in Marketing and Consumer Response

Philipp Iversen

Technical University of Munich

Abstract

Generative AI is revolutionizing the marketing industry by producing high-quality, cost-effective, and time-efficient content. This study investigates the potential of AI-generated music in digital advertising. Two studies, a survey and a real-world A/B test, evaluate different songs on chosen criteria: *Overall, Melodiousness, Creativity, Naturalness, Correctness, and Prompt Following* for the survey, and click-through rates (CTR) for the field experiment. The survey results show that AI-generated music can be comparable in quality to human compositions, even scoring significantly higher in the categories *Prompt Following* and *Melodiousness*. However, AI music showed significantly worse results in the category *Creativity*. The field experiment revealed no statistically significant difference in CTR between advertisements using AI-generated and royalty-free music, demonstrating that AI music can be a good substitute in supporting roles. This research underlines the possibility for AI-generated music to be used in hyper-personalized advertising, while addressing challenges related to perceived creativity and copyright. The findings contribute to understanding AI's disruptive potential in marketing and offer practical insights for integrating AI tools effectively.

Keywords: AI-generated music; music in advertising; AI consumer response

1 Introduction

Generative AI has made significant advancements in recent years. Especially in creative fields it hast become indistinguishable to human made art for most people (Bellaiche et al., 2023). The introduction of ChatGPT, at the end of 2022, was a turning point that is revolutionizing many industries, including marketing. After that there were many new and innovative AI tools with most of them being widely available to the public. In the creative domain, text-to-image tools such as DALL-E, Midjourney, Firefly or Google's Imagen have been able to produce highly effective marketing images. These even outperformed traditional image sources, such as stock images and freelancers, at a fraction of the cost (Hartmann et al., 2024).

With this in mind, and the release of highly sophisticated and easily accessible music generation tools such as Suno AI, Udio AI or StableAudio, the question arises whether AI-generated music can be equally effective for marketing purposes. Music in marketing is very common. According to a study by (Allan, 2008), 94% of prime-television advertisements contained music. And since the global music market has significantly grown since then, it is safe to say that this number is more likely to have increased rather than decreased (IFPI, 2024). But advertisements aren't the only way brands utilize music in their marketing. Most retail stores strategically play music to increase sales as it can make goods and services more appealing which "helps create a bridge between the customer and the product" (Neese, 2015). While the use of music in retail stores is an interesting subject, and the potential of AI-generated music in this context is certainly worth exploring, this paper will focus on its application in online advertisement campaigns.

Music licensing is a complex subject that I won't go too deeply into in this paper. However, it is important to note a few key points. Factors such as the popularity of the song, the duration of use, and the platform on which it will be utilized all influence the cost of a music license. Although it is challenging to generalize licensing fees, they can often represent a significant expense in a company's marketing budget (AudioNetwork, 2023). The rise of AI-generated music offers a promising alternative to help avoid the high costs and legal complexities associated with music licensing. Additionally, AI music can be produced quicker than traditional music. This could, for example, enable hyper-personalized advertisements and in-store music experiences. This has the potential to be highly impactful, as personalized advertisements are often more effective in engaging consumers (Monem, 2021).

When discussing music made by generative AI it is also worth mentioning the complications it brings to the topic of originality and copyright. The questions arises if AI-generated works can be considered truly original. This is particularly important when considering copyright laws, which have traditionally protected works based on their originality and the creator's expression (Fenwick & Jurcys, 2023). Especially, since some of the training data used by the

DOI: https://doi.org/10.5282/jums/v11i1pp181-194

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

most sophisticated AI music generation tools remain undisclosed. Although AI-generated music offers faster and more cost-effective solutions (Lin & Chen, 2024), it challenges these conventional notions of authorship. This development highlights the need for evolving legal frameworks that can balance human-made music with the creations made by AI.

Existing research has mainly focused on the effects of music in marketing in general or the different approaches to generating music using AI, which we will both discuss in the next chapter. Little has been said about the potential effectiveness of AI-generated music in marketing. The role of AI as a disruptive technology in marketing has been explored by various studies, including the 2024 paper by Jochen Hartmann, Yannick Exner, and Samuel Domdey, titled "*Can Generative AI Create Superhuman Visual Marketing Content?*". As this paper aims to address a similar question but for auditory marketing, it draws inspiration from (Hartmann et al., 2024), with the aim of producing equally insightful results.

To achieve this, a two-part survey and a real-life A/B test comparing AI-generated music with traditionally made music will be conducted. The first part of the survey will explore basic consumer attitudes toward AI-generated music. In the second part of the survey participants will listen to short snippets of music, both AI-generated music and music made by a professional musician, and evaluate them based on various criteria. I have selected the metrics outlined by (Chu et al., 2022), but instead of using all nine metrics, I have condensed them into the six most relevant: *Overall, Creativity, Naturalness, Melodiousness, Correctness,* and *Coherence (Prompt following)*. To make it as fair as possible, the various AI tools and the musician will be given the exact same instructions in the form of a text prompt. In line with (Chu et al., 2022), I aim survey novice users without professional musical expertise. Overall, the song generated with Suno AI scored significantly higher values for two out of the six metrics compared to the human-made song, while only scoring significantly worse in one. It also had the highest cumulative mean score, surpassing the human composition as well as the song generated with StableAudio, which ranked last.

For our A/B test, we will compare an AI-generated song with a royalty-free song that shares a similar vibe. To evaluate the effectiveness of these tracks, we will run Facebook and Instagram advertisements, one version with AI-generated music and the other with royalty-free music, under otherwise identical conditions (ceteris paribus). By analyzing Meta's built-in insights, we will compare key metrics, with a focus on the click-through rate (CTR) of both advertisements. The data revealed no significant difference between both advertisements CTR.

The evaluation of these two experiments will provide valuable insights into human perception of AI-generated music. It will also reveal whether low-cost, off-the-shelf AI music generation tools are already sophisticated enough to match or even exceed the quality of traditionally composed music. If so, this paper will showcase yet another example of how AI is revolutionizing the marketing industry. Additionally, we will identify the strengths and weaknesses of AI-generated music, as well as the top-performing tools currently available. From this, best practices can be established. (Davenport et al., 2019).

In the next chapter, I will conduct a comprehensive literature review, starting with the significance of music in marketing. This will be followed by a short look at the history of AI music generation tools and a comparison of different software architectures. I will also go over a few examples of how generative AI is already being used for marketing purposes today. After the literature review, I will present the research design and display the results and analysis of the data obtained through our experiments. Lastly, I will discuss the broader implications of the findings, address potential limitations and suggest future research questions in a future research questions table.

2 Related Work

2.1 Music in Marketing

Brands carefully form their image to resonate with their target audience. To achieve this, every element, from prices, quality and design choices, are carefully considered. Details like color schemes and font selections are carefully chosen to underline the brand's identity. Modern companies focus not just on selling individual products but on creating a cohesive brand experience (Neese, 2015). Therefore, music selection also plays a crucial role in this process. The right soundtrack can underline and enhance what the brand represents, further deepening its connection with consumers (Neese, 2015).

2.1.1 The role of music in advertising

Using music to enhance your marketing efforts is by no means a new idea. In the article (Tom, 1990), music is described as a powerful tool to evoke memories. The paper found that music can significantly help recall advertised products. Interestingly, popular radio hits in their original versions had the lowest recall score, followed by parody songs. Music specifically composed for the advertisement achieved the highest recall rates. Other early studies also examined the influence of music on consumer behavior, showing that different types of music can affect the time spent in a store and the quantity purchased by customers (Keller, 1984). And since music in marketing has proven to be highly effective (Neese, 2015), many studies about its various effects are being conducted to this day.

Before we further look at how music influences consumer behavior, it is important to first consider where and how modern customers listen to music. Two primary distinctions can be made: music played inside physical retail stores or restaurants and music used in advertisements, whether online or on TV. For many decades, most advertisements were shown on TV, but with the recent decline in watch time, online platforms have taken over. On average, internet users now spend roughly 30% more time on social media than watching TV (McKay, 2023). Naturally global spending on social media advertising has significantly increased and in 2023 it

surpassed 207 billion USD (McKay, 2023). One of the key advantages of social media advertising is the ability to get nearly instant customer engagement and feedback. This is particularly valuable when testing campaigns featuring AI-generated music, as the fast feedback allows quick adjustments to an advertisement if necessary.

Ideally both areas of marketing music exposure (in-store and in advertisements) should be used together as suggested by the "Music affect model" by (Kerr & Das, 2014) (Figure 1). The brands advertisements (Online or TV) can positively influence consumer behavior, from the initial exposure to the advertisement to the decision to visit the store. Then the right in-store music can enhance customer behavior during the visit, impact shopping outcomes and boost purchase intentions.

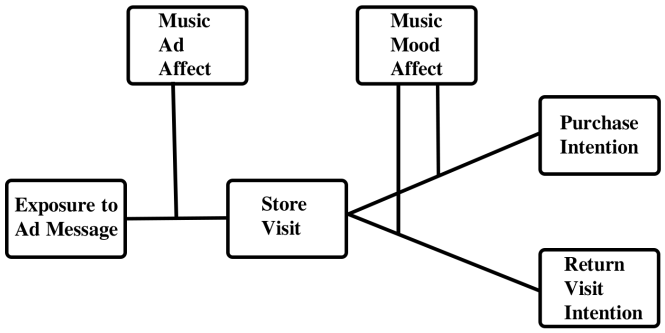

Figure 1: Music affect model (Kerr & Das, 2014)

2.1.2 Psychological and emotional impact of music on consumers

So why is marketing with music so effective? One reason is that humas cannot turn of their hearing ability. "People live in symbiosis with sound and determine dimensions in life through it" (Hultén et al., 2009). Furthermore, humans have a remarkable ability to recall sounds from the past. Music associated with a particular time period can evoke powerful emotions, allowing people to relive memories and, for example, feel youthful again (Hultén et al., 2009). Through cost-effective and rapid music generation AI introduces the capability of creating hyper-personalized music tailored to the demographic profile of individual viewers of advertisements. Another key reason why marketing with music is so effective is the ability of music to influence people's mood. Moods being understood as more muted states of feelings than emotions (Gardner, 1985). And in turn, moods can influence consumer behavior. Many customers tend to make purchases not solely based on a products facts and features, but based on the feelings they experience when thinking of a product (Pham, 1998). This is referred to as the "How do I feel about it?" or *HDIF* heuristic, where positive feelings toward a product make it more appealing and increase the likelihood of purchase. Additionally, a positive mood can help with brand name recognition (Gardner, 1985). Selecting the right song for your marketing efforts can therefore lead to numerous positive outcomes, influencing consumer behavior and brand recall over time.

It has been observed that different types of music can influence different peoples moods in similar ways (Rosenfeld, 1985). Two key factors in any musical piece are tempo and pitch. High-tempo songs, typically classified at 94 BPM or higher, tend to make listeners feel more energized and joyful. On the other hand, lower-tempo songs, around 72 BPM or lower, have a calming effect. This is likely because it aligns with the average resting heart rate of around 72 BPM. Additionally, songs with a higher pitch are generally perceived as happier, whereas lower pitched songs tend to evoke a feeling of sadness (Bruner, 1990). These are just some factors to consider when selecting the right music. Whether for advertisements or in store environments, the right songs can evoke the desired emotional response from customers.

Music plays an important role in marketing, and with an increasing internet usage, its importance is going to grow even further. In a competitive market, music can be a key differentiator, that helps brands in their segmentation and positioning strategy. This enables them to place themselves in a unique market position and reach the right audience. AI music has the ability to create hyper-personalized music tailored to specific consumer groups, or even individual customers, that can provide the competitive edge needed to stand out in today's market.

2.2 Music Generation using AI

In this segment, we will take a brief look at the early attempts and the history of computer-generated music as well as discuss the software architectures and approaches behind different music generation tools. I will also introduce the two state-of-the-art tools that are accessible to everyone and that I will be using for our research in the next chapter.

2.2.1 The History of AI generated Music

The "*Illiac Suite for String Quartet*" is widely regarded as the first song created by a computer (Gage, 2021). In 1957, american composer Lejaren Hiller and mathematician Leonard Isaacson used the *ILLIAC I*, a computer weighing five tons, to conduct experiments in computer-assisted composition. Of course, the methods they used were still very different from the deep learning algorithms we use on today. Hiller and Isaacson used stochastic processes and rule-based algorithms to accept or reject randomly generated rhythms and pitches (Sandred et al., 2009). However, these rules were set by a human, meaning the composer still had a lot of control over the final output. As a result, the system was very limited in its ability to adapt to different musical styles beyond the rules it was programmed with. Nevertheless, the output was a qualitative piece of music created by a computer and therefore bypassing the traditional composition process. This marked the first example of algorithmic thinking in music, which has become more common since (Gage, 2021).

Three decades later in 1985 David Cope developed a computer program, which was initially intended to help him finish a commission. The system developed by Cope is known

as "*Experiments in musical intelligence*" or in short "*EMI*". *EMI* uses a database of songs as input data. It then also uses a rule-based system but focusses on the recombination of musical elements from the database of pre-existing works (Silva, 2003). *EMI* does not rely on randomness in the way the creation of the *Illiac Suite* did, it rather analyzes pieces of music for their structural patterns (melodic, harmonic, and rhythmic) and then recombines these patterns in a logical and stylistically coherent way. *EMI* was developed later with the goal of mimicking musical styles and creating new compositions in the style of specific composers.

2.2.2 Current technology and methods

In the 2010s the emergence of deep learning technologies greatly accelerated the possibilities with AI music generation (Du et al., 2016). Various architectures have contributed to advance the field of AI music. In this section, we will explore some of the recent architectures and discuss their strengths and weaknesses. While the technical details behind these tools are not the main focus of this paper, we will briefly address them for the sake of completeness. Most of the information is taken from these two papers, (Civit et al., 2022) and (Briot, 2020). Please refer to them for a more detailed explanation.

- **Evolutionary algorithms**

  Evolutionary algorithms are modelled after Darwin's theory of evolution. These algorithms start with small musical fragments and modify them through mutation and crossover functions to create new melodies.

  Evolutionary algorithms use principles of music theory as rules. This makes sure that generated melodies fit into the desired output. This method relies on gradual improvement through iteration and is often used together with other methods.

- **Feedforward Networks**

  Feedforward Networks (FF) are a rather simple architecture that serve as a foundation for more complex approaches. As the name suggests, FF networks are unidirectional, which means information is only passed on in one direction through the network, without any loops or recursions.

- **Recurrent Neural Networks**

  Recurrent neural networks (RNNs) on the other hand, are not unidirectional and have also become essential for many music generation architectures. RNNs process sequences by using internal memory to retain information from previous steps, making them particularly suitable for generating musical sequences. The most popular variant is the LSTM (Long Short-Term Memory) networks. It can capture long-term dependencies, allowing it to generate high quality compositions over time.

- **Generative Adversarial Networks**

  Generative Adversarial Networks (GANs) consist of two competing networks: a generator and a discriminator. The generator creates an output, in our case a song, that the discriminator then takes as input to decide how realistic it is. This creates a feedback loop in which the generator improves its ability to create music that the discriminator cannot easily distinguish from human made compositions. GANs are often combined with other architectures like feedforward or recurrent networks.

- **Variational Autoencoder**

  Variational Autoencoder (VAE) maps input music into a latent space, capturing essential features while leaving out unnecessary details. The decoder part then uses a distribution to map each point back to the input space which generates new pieces. VAE's are particularly effective for exploring creative variations within musical data by modifying the latent space. VAEs are not only used for composition but also work well audio compression.

- **Transformer Model**

  In recent years, Transformer Networks have become more popular due to their ability to model long-term dependencies in sequential data, such as music. Unlike RNNs, which process data sequentially, transformers use attention mechanisms to consider all parts of a sequence at once. This makes them particularly useful for generating complex compositions.

- **Latent Diffusion Model**

  The newest addition to the list of music generation architectures is latent diffusion. The system uses a VAE, which compresses raw audio into a lower-dimensional latent space. Then it uses a diffusion process to refine and progressively generate audio content. The diffusion model operates iteratively, adding small modifications to the latent representation until the desired audio output is achieved (Evans et al., 2024). The use of latent diffusion makes the system computationally efficient, which is beneficial for generating long-form audio at high quality.

Together, these architectures demonstrate the fast evolution of AI for music generation. As these technologies continue to develop, AI is going to become an even more powerful tool in the world of music creation.

2.2.3 Suno AI & StableAudio

In the final part of this chapter, we will explore the state-of-the-art tools which will be used for our research in the following chapters. The two tools we will focus on are Suno AI and StableAdio. They both launched in late 2023, which makes them among the first, easy-to-use online platforms

that allowed users to create full length songs from a single text prompt.

As of September 2024, both music generation tools offer a similar free plan and pricing strategy. The free plan limits the number of songs that can be generated within a certain period. The paid plans, ranging from 10 Euro to 90 Euro per month for StableAudio's max tier, increase the number of songs that can be generated. They also unlock some additional features such as commercial use, priority in the generation queue for a faster song creation process and more simultaneous generation jobs. Furthermore, Suno and StableAudio allow you to upload audio files as reference. However, none of the mentioned tools have options for further customizing or modifying the generated songs (StableAudio.com, 2024; Suno.com, 2024).

StableAudio 2.0, which is Stability.ai's newest version claims to have been "exclusively trained on a licensed dataset from the *AudioSparx* music library, honoring opt-out requests and ensuring fair compensation for creators" (Stability.ai, 2024). Suno, on the other hand, has kept their exact training data a secret. This raises copyright issues which has already led to a law suit by the Recording Industry Association of America (RIAA) (Rogès, 2024).

What is known about their models, however, is that they not only use songs but also voice recordings to train their AI. More specifically Suno mainly utilizes a transformer networks architecture to generate music (NoPriors-Podcast, 2024). StableAudio, on the other hand, uses a latent diffusion model. Its architecture is made up of a autoencoder that compresses the raw audio data and a diffusion transformer (DiT) that "refines random noise into structured data incrementally, identifying intricate patterns and relationships" (Stability.ai, 2024).

2.3 AI-generated Music in Marketing

In this chapter we will take a look at how AI is already being used in marketing today. There have already been many successful advertisement campaigns using AI. For example, in 2023 Coca-Cola held an art contest in which they invited digital artists to create Coca-Cola themed artworks. This was done with the help of a custom-built image generation platform based on OpenAI's GPT-4 and Dall-E (Coca-Cola, 2023). For Nike's 50th anniversary their campaign "Never Done Evolving" included two digital models of the tennis star Serena Williams playing each other. The innovative part was, that each model was trained with real data gathered from her real matches using AI. The final was then broadcasted on YouTube with 1.7 millions viewers, a significant increase of views compared to Nikes other content (WPP, 2022). Or in 2019 JPMorgan Chase partnered with Persado, an AI Text tool which focusses on generating texts for marketing purposes, to rewrite their ad copy. This increased JPMorgan Chase's CTR by 450% (Persado, 2019). These were just a few examples of how AI is already being used in marketing today.

But what about advertisement campaigns utilizing AI-generated music specifically? In early 2024 the American sea food chain Red Lobster created a playlist on YouTube with 30 songs about their highly popular Cheddar Bay Biscuits. All the songs in this advertisement campaign were AI generated and covered a wide variety of genres (Nelson, 2024). Several of the songs have become some of the most watched videos on the channel, with the majority of comments praising the biscuits or the AI-generated music for their quality and innovativeness. Overall, the advertisement received a highly positive response from its audience, leaving a positive impact. With modern AI music tools such as the ones discussed in the previous chapter, it is safe to assume that the advertisement campaign was inexpensive yet, judging by the consumer response, highly effective.

In my research I was unable to find more examples of established brands using AI-generated music in advertisements. This might be due to easily accessible music generation tools still being very new in their availability to the public. There is also the possibility that AI-generated music has been used in advertisements discreetly, as the general public still has a bias against AI music, and publicly acknowledging its use might attract unwanted criticism (Zenieris, 2023). Furthermore, some studies suggest that AI-generated music is now often indistinguishable from human composed music (Highams & Olszewska, 2023). This means there is a possibility that AI music is already in wider use than we realize. What is known, however, is that content creators are already using AI-generated music to produce content on platforms like YouTube, Facebook and TikTok. This allows them to avoid copyright issues and generate revenue more freely (Soundful, 2024).

The adoption of easy to use off the shelf AI music tools is still in its early phase. The examples presented have demonstrated the potential of AI being used in marketing campaigns. To further assess the effectiveness of AI-generated music we will gather and analyze data from real-world studies in the next chapter.

2.4 Hypotheses

Building on the findings of this chapter, this study aims to bridge the gap between existing research on AI-generated music and its practical applications for marketing. Given the limited research on how AI-generated music might perform against traditional compositions in digital advertisements, I came up with four hypotheses:

- **H1**: *Casual music listeners will not significantly differentiate between AI-generated music and human-composed music in terms of overall quality.*
- **H2**: *The choice of the AI models architecture will impact the overall satisfaction of AI-generated music by casual listeners.*
- **H3**: *A casual listeners attitude towards AI technology will positively correlate with their ratings of AI-generated music for marketing purposes.*

- **H4**: *AI-generated music will achieve a higher click-through rate (CTR) than royalty-free music when used in identical advertisements on social media.*

3 Empirical Investigation

3.1 Research Design

Two studies were conducted to obtain empirical data to answer the question if AI generated music can be as effective, if not more effective than traditionally composed music in advertising. Through rejecting or failing to reject our previously outlined hypotheses the research questions, stated in the following table (Table 1), will be answered.

Table 1: Overview of the two studies

	Study 1	**Study 2**
Research Question	How do casual music listeners perceive AI-generated music for marketing purposes compared to traditionally produced music, and does the choice of AI model architecture influence these perceptions?	Can AI-generated music outperform royalty-free music in digital advertising by achieving a higher click-through rate when used in identical ads on social media platforms?
Dependent Variables	Overall, Creativity, Naturalness, Melodiousness, Correctness (Chu et al., 2022) Prompt Following	CTR (Click-through rate)
Setup	Online Survey (Lab setting) • 2 AI-generated songs (Suno AI based on transformer models and StableAudio based on transformer diffusion models) • 1 song made by a human producer • 41 participants	Online Advertisement (Field experiment, A/B Testing) • 1 ad with AI generated song (Suno AI) • 1 identical ad with Royalty free music • Facebook Advertisement: 10 days for 200€ (10€ per day per advertisement) • 52680 total impressions

3.2 Online Survey (Lab Setting)

The initial study is made up of a two-part survey. It begins with collecting general information about the participant, such as age, gender and their general attitude towards AI technology and AI-generated music. All questions in the first and second part of the survey, except the questions about age and gender, are answered on a 5-point Likert scale, where 1 indicates strong disagreement, and 5 strong agreement with the presented statements.

In the second half of the survey the respondent listens to three song snippets. Two songs are AI-generated and one was made by a human producer with an expertise in commercial music composition. For the AI-generated songs, I utilized Suno AI, which primarily employs transformer models, and StableAudio, which is based on diffusion models. By comparing these two different software architectures to a human producer, the goal is to determine which approach performs best and whether the choice of model architecture significantly impacts the perceived quality of the music. The participants were not informed about whether a song was AI-generated or not. In fact, no indication was given regarding the origin of the songs. This approach was intended to prevent listeners from searching for signs of AI involvement, which could have distracted them from providing unbiased ratings.

To make the comparison as fair as possible, the exact same prompt was given to both AI tools as well as the human producer with no further instructions or communication, except that the song must be an instrumental. The prompt provided was the following: “Happy, upbeat, summer hip-hop beat for a drink commercial”. This prompt was intentionally formulated in a way that includes an emotion (happy) and a musical genre (Hip-Hop), ensuring that all the generated songs would follow a similar direction. Including the context of a drink commercial was also crucial, as this would of course be known in a typical advertisement campaign briefing. Additionally, I introduced a subtle contradiction by combining the description “upbeat” with “hip-hop beat”, two elements that can have contrasting sonic qualities. This added some room for creative interpretation. For each prompt Suno AI automatically generates two songs. I selected the one that sounded best to me, as this is how someone would realistically choose AI-generated music for an advertisement.

Each snippet is exactly 30 seconds long. This gives the listener enough time to get a good idea of the song without risking listening fatigue. For all three songs the first 30 seconds were chosen. After each song snipped there are six questions to assess the participants impressions. The questions are derived from (Chu et al., 2022), but instead of the nine criteria outlined in that paper I have condensed it down to the five most relevant plus a sixth criteria which assesses how closely the song aligns with the given prompt. The six criteria are:

- **Overall**: What is your overall satisfaction with the song?
- **Creativity**: Is the song novel, valuable, and original?
- **Naturalness**: Does the piece sound like an expressive human performance?
- **Melodiousness**: How musical and harmonious is the song?
- **Correctness**: Does the song play with any technical glitches?
- **Prompt Following**: How well does the song fit the given prompt?

Consistent with the approach of (Chu et al., 2022) the survey targeted casual music listeners without any formal background in music. This was done to get the most authentic and natural ratings, simulating a real-world scenario for advertisements. A clear definition for every criteria as well as the prompt, used to create the songs, was provided.

3.3 Online Advertisement (Field Experiment)

The second part of the research is a field experiment, conducted by running paid advertisements on social media platforms. I created two identical 15 second animated clips that advertise *Imparat*[1] paint for your home. I partnered with IMparat Farbwerk GmbH & Co KG to obtain the funding for the real-world A/B test and to help them kick of their social media advertisement campaign. I added music generated by Suno AI to one video and a royalty-free song from Pixabay to the other. I chose Suno AI as my AI song generator because the results of my online survey indicated early on that it would outperform the song generated by the diffusion model of StableAudio. To make it as fair as possible I created a simple prompt that could also be used as a search term in the Pixabay audio library. The prompt I used was "Chill melodic lofi beat". I again chose the song from the two generated by Suno AI that sounded best to me. A similar approach was used when selecting the royalty free song. Within the first page of the search results of my given prompt I selected the song I found most suitable.

The advertisements were then published through Meta's Ad Center, which showed them on Facebook and Instagram. I chose Meta because it offers an uncomplicated way to publish advertisements with a low daily budget. Additionally, it offers insights into the advertisement's performance providing crucial data such as total impressions and link clicks, which we need to determine the CTR. Both advertisements were uploaded at the same time, with a runtime of 10 days (27.10.2024 – 05.11.2024) and a daily budget of 10 Euro. As mentioned by (Hartmann et al., 2024) this type of real world A/B test aims "to increase the ecological validity of our findings".

[1] Imparat is a paint company based in Hamburg, Germany

4 Results

4.1 Participants

For the online survey we had 41 participants of which 55.6% identified as female, 37.2% as male and 7% as diverse. With around 60%, most of the respondents were between 21 – 30 years old. 18.5% were 46 – 60 years old with the rest being relatively evenly distributed among the remaining age groups. All the participants are based in Germany.

Our field experiment generated a total of 52680 impressions, with approximately 60% of the reached accounts being male. The age distribution was relatively balanced across all groups between 25 and 64 years, accounting for around 22% each. The remaining 12% of the profiles reached were in the age range of 65+. The advertisement campaign targeted geolocations where IMparat Farbwerk GmbH & Co KG has a physical presence, resulting in all reached accounts being based in Germany as well.

4.2 Results – Online Survey

Overall, our participants have a positive attitude towards AI technology. With an average of 3.49 on a 5-point Likert scale where 1 indicated a negative, and 5 a positive view on AI technology. At first this might seem suspiciously high when considering that according to (Schneider, 2021) around 40% of Germans are skeptical towards AI technology. Taking into account the younger demographic of our survey, the higher acceptance rate makes sense since only 19% of the AI skeptics in Germany are between the ages of 18 and 34. This age group makes up 24% of the total population of Germany (Schneider, 2021). With an average of 3.39 the origin of a song also seems to be important to our respondents. The question if AI songs are still considered art was more divisive. With an average score of 2.61, participants are roughly balanced in their disagreement and agreement with that statement.

Which song performed the best? To answer this question, we took the average of all six song categories and plotted them against each other as seen in Figure 2. Song 1, which was generated by Suno AI achieved the highest score in the *Overall* category with an average of 3.42. Followed by the human composed song with 3.27 and the song generated by StableAudio with a score of 2.73. For the category *Creativity*, *Naturalness* and *Correctness* the human-made song scored the highest means with 3.54, 3.15 and 4.32. Song 1 got the highest average score for *Melodiousness* and *Prompt following* with 3.98 and 4.05.

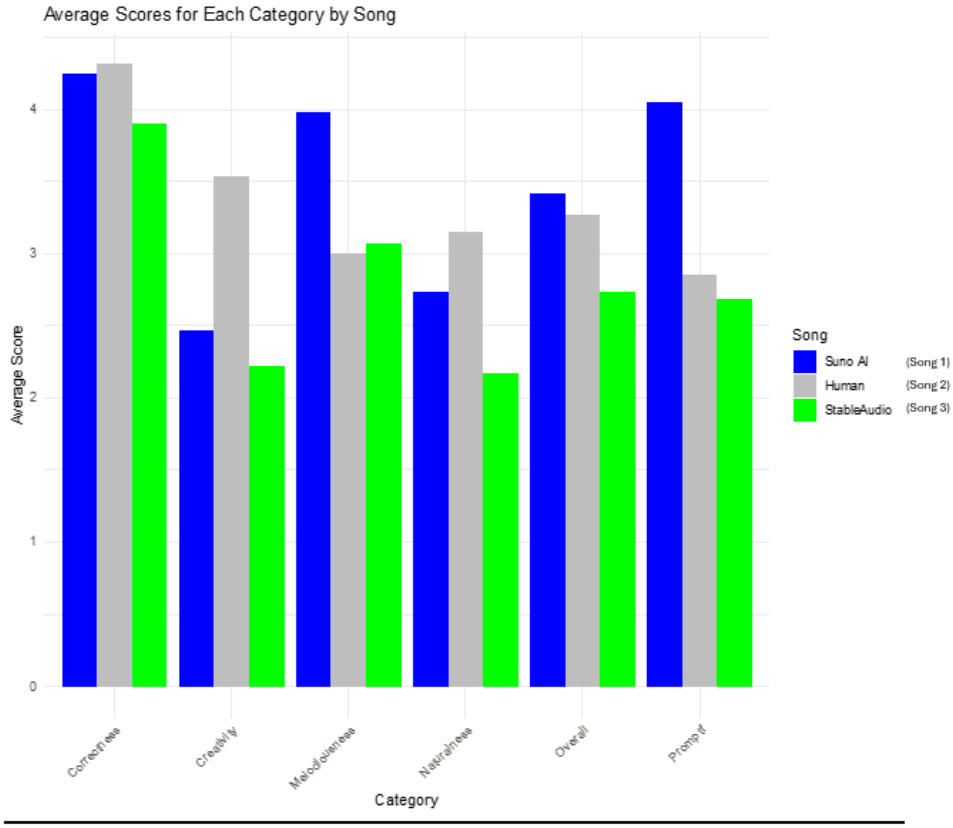

Figure 2: Bar chart of the means of each song category

To investigate if the differences in means are significant, I ran an OLS regression on all six song related categories and used Song 2 from my survey, which was the human-made song, as a baseline. The results are displayed in Table 2.

Table 2: OLS regression results on the impression of AI music compared to human made music in an advertisement context

Human made song vs AI generated songs						
	Dependent variable:					
	Overall	**Creativity**	**Naturalness**	**Melodiousness**	**Correctness**	**Promptfollowing**
	(1)	(2)	(3)	(4)	(5)	(6)
Suno AI	0.146	-1.073^{***}	−0.415	0.976^{***}	−0.073	1.195^{***}
	(0.261)	(0.248)	(0.274)	(0.228)	(0.235)	(0.233)
StableAudio	-0.537^{**}	-1.317^{***}	-0.976^{***}	0.073	-0.415^{*}	−0.171
	(0.261)	(0.248)	(0.274)	(0.228)	(0.235)	(0.233)
Constant	3.268^{***}	3.537^{***}	3.146^{***}	3.000^{***}	4.317^{***}	2.854^{***}
	(0.185)	(0.175)	(0.194)	(0.161)	(0.166)	(0.165)
Observations	123	123	123	123	123	123
Within R^2	0.0437	0.1969	0.081	0.1453	0.0125	0.2408
Observations	123	123	123	123	123	123
R^2	0.059	0.210	0.096	0.159	0.029	0.253
Adjusted R^2	0.044	0.197	0.081	0.145	0.012	0.241

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Most notably in the category *Creativity* both Suno AI ($\beta_{Creativity} = -1.073$, $p < 0.01$) and StableAudio ($\beta_{Creativity} = -1.317$, $p < 0.01$) had highly significantly worse outcomes compared to the human-made baseline. Across all six categories StableAudio had significantly lower values in four of them. Besides the previously mentioned creativity score, it had a significantly lower score for *Overall* ($\beta_{Overall} = -0.537$, $p < 0.05$), a highly significantly lower score for *Naturalness* ($\beta_{Naturalness} = -0.976$, $p < 0.01$) and a marginally significantly lower score for *Correctness* ($\beta_{Correctness} = -0.415$, $p < 0.1$). On the contrary, Suno AI did not perform significantly worse in any category, except for the already mentioned *Creativity* attribute. It was able to significantly outperform the human-made song in two categories: *Melodiousness* ($\beta_{Melodiousness} = 0.976$, $p < 0.01$) and *Prompt Following* ($\beta_{Promptfollowing} = 1.195$, $p < 0.01$). It is important to note that, although the differences in values are statistically significant, the real-world differences may be minimal. Since the results are based on a 5-point Likert scale, some of the observed values are only slightly above or below the baseline when examining the actual scores.

Combining all the means for each song, Song 1, which was generated using Suno AI, achieved the highest score of 20.88. Closely followed by the human made song with 20.12. StableAudio had the lowest aggregate score with 16.78.

To examine the influence of selected variables on the *Overall* metric for each song, I conducted regression analyses (Table 3). In the first model, I used only the participants attitudes towards AI as independent variables: whether they like AI, whether the origin of a song is important to them, and whether they consider AI-generated music to be art. In the second model, I also included the song metrics: Creativity, *Melodiousness*, *Correctness*, *Naturalness*, and *Prompt Following* as independent variables.

Table 3: Regression on the *Overall* metric of each song

		Suno AI		**Human song**		**StableAudio**	
	Variable	ovReg1	ovReg11	ovReg2	ovReg22	ovReg3	ovReg33
1	Like_AI	−0.0231 (0.85129)	0.22398 (0.12032)	−0.07057 (0.72902)	−0.01708 (0.8837)	0.1528 (0.25881)	0.23643 (**0.04345**)
2	Song_Origin	0.03884 (0.76255)	−0.02569 (0.86066)	0.29121 (0.17559)	0.05589 (0.67965)	−0.00647 (0.96315)	−0.05557 (0.62501)
3	AI_is_Art	0.40536 (**0.00173**)	0.21116 (0.19765)	0.01815 (0.92759)	−0.24191 (**0.04661**)	0.42264 (**0.00257**)	0.07538 (0.51923)
4	Creativity		0.24313 (0.38425)		0.01366 (0.95699)		0.25075 (0.15242)
5	Naturalness		0.1184 (0.56624)		0.39057 (**0.0419**)		0.15019 (0.41823)
6	Melodiousness		0.27426 (0.25528)		0.68432 (**0.00097**)		0.42172 (**0.0085**)
7	Correctness		−0.0595 (0.7187)		−0.21047 (0.14952)		−0.02534 (0.84768)
8	Promptf		−0.17305 (0.47609)		0.23489 (**0.06585**)		0.14736 (0.32751)

The table shows that a participant's general attitude toward AI technology only had a significant positive effect on the overall perception of the song generated by StableAudio when the regression included additional song metrics ($\beta_{Like_AI} = 0.236$, $p < 0.05$). The importance participants placed on the origin of a song did not significantly affect their overall perception of any song. Agreeing strongly with the statement that "AI music still counts as art" appears to have a positive impact on the *Overall* metric for both AI-generated songs: Suno AI ($\beta_{AI_is_Art} = 0.405$, $p < 0.01$) and StableAudio ($\beta_{AI_is_Art} = 0.423$, $p < 0.01$). For the human generated song, however, this belief had a significantly negative effect when all independent variables were included in the regression ($\beta_{AI_is_Art} = -0.242$, $p < 0.05$). Among all song-related metrics, *Melodiousness* had the highest positive effect on the overall perception of the song. For two out of the three songs, it had a highly significant impact: the human-made song ($\beta_{Melodiousness} = 0.684$, $p < 0.01$) and the StableAudio song ($\beta_{Melodiousness} = 0.422$, $p < 0.01$). This aligns with the findings by (Chu et al., 2022), where *Melodiousness* also had the greatest impact on a songs overall impression.

4.3 Results – Online Advertisement

Analyzing the data from my social media ad campaign was much simpler. The advertisement video with the AI-generated music by Suno AI had a total of 23055 impression and 206 link-clicks after 10 days. The advertisement with the royalty-free music had 29625 impressions with 246 link-clicks during the same time period. This leads to a CTR (click-through-rate) of 0.894% for the advertisement with the AI music and 0.830% for the advertisement with the royalty-free music. To investigate the significance of the click-through rates (CTR) between the campaigns, a chi-squared test was conducted (Table 4). The null hypothesis for this test assumes that the CTR's for both campaigns are statistically identical, with any observed differences being due to random variation.

The Chi-squared test concluded that the difference in CTR between the two ads is not significant. This is indicated by the relatively low chi-squared value (0.5356) in line 5 of Table 4 and the p-value (0.4643) in line 7, which is above the typical threshold of 0.05. We therefore cannot reject the null hypothesis and conclude that neither the advertisement with the AI or the royalty free music performed significantly better or worse than the other.

Table 4: Chi-squared test results for ad CTR comparison

	Metric	Suno_AI	Royalty_Free	Test_Results
1	Click–through Rate (CTR)	0.89%	0.83%	
2	Observed Clicks	206	246	
3	Total Impressions	23055	29625	
4	Difference in CTR			0.06%
5	Chi–squared Value			0.5356
6	Degrees of Freedom			1
7	p–value			0.4643
8	95% Confidence Interval (Lower)			−0.10%
9	95% Confidence Interval (Upper)			0.23%

4.4 Analysis of Findings for the Research Questions & Hypotheses

After analyzing the data, we can now answer the research questions outlined in Table 1 and evaluate the hypotheses presented in chapter 2.4.. The results do not support H1, which means that casual music listeners do significantly differentiate between AI-generated and human produced music in terms of perceived overall quality. As shown in Table 2, the song generated by Suno AI scored significantly higher than the human composed song in two out of six categories: *Melodiousness* and *Prompt Following*. Table 4 revealed that *Melodiousness* has the largest impact on the *Overall* metric. Although Suno AI's song achieved a numerically higher mean in the *Overall* category than the human-composed song, this difference was not statistically significant. However, in advertising contexts, the *Prompt Following* metric is crucial, as it indicates how well the music aligns with specific instructions. Notably, Suno AI's song only scored significantly lower than the human song in the *Creativity* category. On the contrary, the song created by StableAudio scored significantly lower than the human composed song in four out of the six categories. Only *Melodiousness* showed a numerically higher score. Since one AI-generated song performed better than the human composed song and the other performed worse, we must reject H1, as there is clear evidence that listeners differentiate the perceived overall quality of AI-generated and human composed music. These findings lead us right into Hypothesis 2, which suggests that the choice of AI model affects how people judge the quality of music. Since the song generated by Suno AI significantly outperformed the song made with StableAudio, we fail to reject H2.

Our third hypothesis (H3) suggests that a positive attitude towards AI technology correlates with higher ratings for AI-generated songs in a marketing context. Table 3 shows the results from four regressions conducted on the two AI songs. Of these, three show a significant positive effect in at least one of the three categories related to AI technology attitude *(Like_AI, Song_Origin* and *AI_is_Art)*. The category *AI_is_Art* had a significant positive effect on both AI songs but only for the regressions where only AI related variables were included. Additionally, for the AI song made with StableAudio, the category *Like_AI* had a positive effect on the song's overall score. This time only in the regression that included all other variables. Interestingly, for the human composed song, *AI_is_Art* had a negative effect on the *Overall* rating. All in all, the data delivers some indicators supporting H3, but the data is not entirely consistent, and the results should be interpreted with this in mind.

The fourth hypothesis proposed that AI-generated music, when used in an otherwise identical social media advertisement, would achieve a higher click-through rate (CTR) than human-made, royalty-free music. As shown in Table 4, the advertisement with the music from Suno AI did have a numerically higher CTR than the advertisement with royalty-free music (0.89% vs. 0.83%). However, the chi-squared test revealed that this difference is not statistically significant. We can therefore reject the H4 hypothesis. These findings provide clear answers to the research questions presented in Table 1, contributing to our understanding of the potential of AI-generated music in marketing.

5 Discussion

5.1 Summary

High-quality AI-generated music is not just a potential technology of the future, but a reality today (Lin & Chen, 2024). Some studies suggest that AI-generated music is already indistinguishable from human-made compositions (Chu et al., 2022; Highams & Olszewska, 2023). However, other studies also show a persisting bias against AI music (Zenieris, 2023). This is an obstacle AI-generated art in general still needs to overcome. Other than its quality, there are other compelling reasons for using AI music, such as significant advantages in production time and cost compared to human-made music.

In our survey, the AI-generated songs took less than two minutes to create, and the cost was effectively zero, as we used a free plan for non-commercial purposes. In contrast, the song created by a human producer took approximately three days to make, not because it required that much time to be produced, but because of limited human availability. While, due to a personal connection, the producer volunteered for this project, he provided a hypothetical invoice of 200 Euro, based on four hours of work at a standard rate

of 50 Euro per hour. Given this producer's relatively young age and freelance career, this cost estimate can be considered conservative.

For our field experiment, the costs were slightly different. Because the AI-generated music was used in a commercial setting it required a paid subscription to Suno AI. I chose the lowest tier, which cost 10 Euro per month. This would have allowed me to generate up to 500 commercially usable songs. Since I only needed one song, I calculated the effective cost to be 10 Euro. Although the royalty-free song used in the study was free, finding a suitable royalty-free song that met all criteria was more challenging than I had anticipated, with limited options available.

The data from our questionnaire indicates that a human producer may not always produce a song that perfectly aligns with a given description. In the *Prompt Following* category, Suno AI achieved a significantly higher mean score than the human produced song. Even StableAudio, which performed significantly worse overall, only showed a numerically lower score in this category. This suggests that while human producers bring a unique style, influenced by personal taste and experience, this can sometimes get in the way of a precisely chosen description. On the other hand, when a client is not sure what exactly they are looking for, this can be a great advantage. When the desired attributes are clearly defined, however, AI can generate a song that meets these specifications with equal, if not greater, accuracy.

Additionally, AI enables rapid iterations. If the generated music does not fully meet expectations, slight adjustments to the prompt can produce a new version within minutes. This capability introduces new possibilities for advertising. Multiple versions or styles of a song could be generated quickly and at a low cost. Advertisers could then conduct real-world A/B tests and evaluate which version performs best with their target audience, always choosing the most effective option (Schwartz et al., 2017). This rapid feedback loop, directly involving the target audience, can help increase effectiveness and engagement (Fletcher, 2024).

The already discussed low cost, rapid creation, and high flexibility of AI-generated music open up possibilities for hyper-personalized advertising, a method that has been shown to be more effective than traditional advertising (Monem, 2021). With high-quality, user-friendly AI music tools, it is possible that future advertisements could feature music customized for individual user preferences. These song customizations could be based on data such as age, location, language, and browsing behavior. Such a level of personalization would be impractical, if not impossible, without the efficiency of AI tools.

There are, however, also potential downsides to using AI-generated music. On a technical level AI is currently limited in its ability to generate highly specific or niche requests, that require a level of creativity which currently only humans have. This limitation was supported by the results of Study 1, where the *Creativity* category was the only one receiving significantly lower ratings for both AI models compared to the human composed song. Although we lack data on whether participants recognized the AI-generated songs, and thus rated them lower in creativity, it is possible.

Moreover, there remains a general bias against AI-generated music (Zenieris, 2023), which could be a risks for companies using this technology. If audiences perceive AI-generated music negatively, it could impact the brands image and reduce the effectiveness of their marketing campaigns. This bias is likely due to ethical concerns, such as copyright issues related to the training data used for AI models. There is also the issue that AI-produced music could closely resemble works by known artists without paying them royalties. Addressing these concerns will be essential to achieving broader acceptance and integration of AI-generated music in commercial settings.

5.2 Implications for Marketing

AI-generated music has emerged as a valuable tool in the marketing industry, offering both advantages and disadvantages. For smaller companies with limited budgets, AI-generated music can help level the playing field (Acar & Gvirtz, 2024) by providing high-quality compositions at a fraction of the cost of a traditionally licensed song. This is particularly beneficial in competitive markets where established corporations have much bigger budgets for their marketing strategies. The data from our second study highlights that AI-generated music, produced using readily available tools, has reached a high level of sophistication. This allows it to be seamlessly integrated as background music, enhancing other elements of the marketing mix.

On the other hand, the current state of AI technology has its limitations when producing music that aims to have a lasting cultural or emotional impact. AI lacks the true creativity required to compose iconic songs that remain memorable and elevate a brand's identity. As of six months after its launch, Suno AI had attracted over 12 million users (Tencer, 2024). Nevertheless, only a single fully AI-generated song has made it into the German music charts[2] (Kramer, 2024), and not a single AI song has achieved this in the U.S. This suggests that while AI is a powerful tool for specific applications, it still needs to demonstrate its potential for creating hit songs capable of deeply engaging audiences.

5.3 Limitations of the study and suggestions for future research

First, I will address the technical limitations of my research. The questionnaire included 41 participants, with over 65% being 30 years old or younger. This demographic distribution could have introduced a sample bias. The results could be influenced disproportionately by the majority of younger participants. Additionally, the sample size prevented meaningful analysis of data across different age groups, as there would not be enough data points within each

[2] In August of 2024 the song "Verknallt in einen Talahon" by Butterbro debuted in the top 50 of the German charts. The song, except for the lyrics, was entirely AI generated using Udio AI

Table 5: Future research question table

Area of focus	Research question
Demographic influence	How do different age groups perceive AI music?
Platform-Specific effectiveness	How does AI generated music perform in advertisement on platforms such as YouTube and TikTok?
AI vs professional music productions	How does AI-generated music compare to tracks produced by high-budget professional compositions?
Long-Term impact	Can AI-generated music improve long-term marketing outcomes, such as brand recall and customer loyalty?
Consumer bias and perception	How does consumer bias against AI-generated music impact its effectiveness in marketing campaigns?
Hyper-Personalization potential	How effective is hyper-personalized AI-generated music in marketing?
Ad placement context	How does the context (e.g. type of product, target audience) affect the success of AI-generated music?
Legal and ethical considerations	How do copyright and originality issues affect the adoption of AI-generated music in commercial use?
General AI music applications	Beyond marketing, how can AI-generated music be effectively used in industries like film, gaming, and education?

segment for a reliable differentiation. Future research could address this by targeting a larger sample.

Secondly, in our field experiment we were able to obtain a large sample size but were limited to comparing only one AI tool against one royalty-free song. While Suno AI was able to outperform the human composition and the song made by StableAudio in our first study, there are many different AI music tools, each with unique strengths and weaknesses. A broader investigation including multiple AI tools would provide a more comprehensive understanding of the current state of AI-generated music and how it compares to human-made compositions.

Thirdly, according to (Jones, 2021), approximately 85% of Facebook users use the platform without sound. Given that the majority of our impressions were gathered on Facebook, and no data was available whether the advertisements were viewed with or without sound, it is unclear how this might have influenced the evaluation of our advertisements. Future research could make use of other media platforms such as YouTube or TikTok, where music and sound is a core feature of the platform, resulting in less users viewing the advertisements without sound (Jones, 2021).

Additional limitations, outlined in the research by (Hartmann et al., 2024), also apply to this study. Our research focused on the qualitative perception of AI-generated music, measured through categories in Study 1, and the immediate user response in Study 2. Future studies could expand on this by exploring its effectiveness in different contexts and examining additional marketing objectives, such as brand recall or purchase intention. Lastly, we also only compared AI-generated music against a lower-budget freelance producer. Additional studies could investigate how AI music holds up against big budget productions.

These considerations provide a strong foundation for future research. A future research question table is provided with Table 5.

6 Conclusion

The rapid evolution of generative AI has made AI-generated music a promising tool for marketing. This study aimed to explore the effectiveness of AI-generated music in a marketing context. This was done by assessing its perception among casual listeners and investigating the performance in a real-world advertisement campaign. The results reveal both potential and some limitations, suggesting that AI music can be very effective, when used in the right context.

AI-generated music showed its ability to outperform human composed songs in several areas, most notably in the categories *Melodiousness* and *Prompt Following*. With tools such as Suno AI and StableAudio songs could be produced in a fraction of the time and at a significantly lower cost compared to on demand made human music. These attributes make AI-generated music particularly appealing for companies seeking fast, flexible, and budget-friendly solutions for their marketing campaigns. However, the study also highlights limitations of AI-generated music. The data of our study shows that the creative quality of AI-generated music, is still significantly lower than that of human composed music. This finding underlines a fundamental challenge for AI in creative industries: the ability to produce compositions that evoke the same level of passion and originality as human creations. While AI excels in following technical prompts and generating music that fits predefined criteria, it struggles to create truly innovative or iconic tracks that have the potential to contribute to long-term brand identity. Our field experiment suggests that while AI-generated music may not significantly outperform traditional options in terms of CTR, it remains a viable alternative, particularly when budget and time constraints are critical factors.

In the future, advancements in AI technology will likely address many of the current limitations. Developing models that can further enhance creativity and emotional resonance could bridge the gap between AI-generated and human composed music. Additionally, larger, more diverse studies are needed to examine how preference for AI-generated music varies across different demographic groups and cultural contexts. This research serves as an initial step toward understanding the potential of AI music in marketing.

References

Acar, O., & Gvirtz, A. (2024). GenAI Can Help Small Companies Level the Playing Field. *Harvard Business Review*. Retrieved March 20, 2026, from https://hbr.org/2024/02/genai-can-help-small-companies-level-the-playing-field

Allan, D. (2008). A Content Analysis of Music Placement in Prime-Time Television Advertising. *Journal of Advertising Research*.

AudioNetwork. (2023). Trends in Music Licensing for TV Advertising. *AudioNetwork*. Retrieved March 20, 2026, from https://blog.audionetwork.com/the-edit/advertising/trends-in-music-licensing

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., & Seli, P. (2023). Humans versus AI: Whether and Why We Prefer Human-Created Compared to AI-Created Artwork. *Cogn Res Princ Implic*, *8*(1), 42. https://doi.org/10.1186/s41235-023-00499-6

Briot, J.-P. (2020). From Artificial Neural Networks to Deep Learning for Music Generation: History, Concepts and Trends. *Neural Computing and Applications*, *33*(1), 39–65. https://doi.org/10.1007/s00521-020-05399-0

Bruner, G. C. (1990). Music, Mood, and Marketing. *Journal of Marketing*.

Chu, H., Kim, J., Kim, S., Lim, H., Lee, H., Jin, S., Lee, J., Kim, T., & Ko, S. (2022). An Empirical Study on How People Perceive AI-Generated Music. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. (2022). A Systematic Review of Artificial Intelligence-Based Music Generation: Scope, Applications, and Future Trends. *Expert Systems with Applications*, *209*. https://doi.org/10.1016/j.eswa.2022.118190

Coca-Cola. (2023). Coca-Cola Invites Digital Artists to 'Create Real Magic' Using New AI Platform. Retrieved March 20, 2026, from https://www.coca-colacompany.com/media-center/coca-cola-invites-digital-artists-to-create-real-magic-using-new-ai-platform

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2019). How Artificial Intelligence Will Change the Future of Marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Du, X., Cai, Y., Wang, S., & Zhang, L. (2016). Overview of Deep Learning.

Evans, Z., Carr, C., Taylor, J., Hawley, S. H., & Pons, J. (2024). Fast Timing-Conditioned Latent Audio Diffusion. *International Conference on Machine Learning*.

Fenwick, M., & Jurcys, P. (2023). Originality and the Future of Copyright in an Age of Generative AI. *Computer Law & Security Review*, *51*. https://doi.org/10.1016/j.clsr.2023.105892

Fletcher, C. (2024). Feedback Loops: How Digital Marketers Can Continuously Improve With Customer Insights.

Gage, J. (2021). 'He Touched a Nerve': How the First Piece of AI Music Was Born in 1956. *The Guardian*. Retrieved March 20, 2026, from https://www.theguardian.com/music/2021/dec/07/he-touched-a-nerve-how-the-first-piece-of-ai-music-was-born-in-1956

Gardner, M. (1985). Mood States and Consumer Behavior: A Critical Review. *Journal of Consumer Research*, *12*(3), 281–300.

Hartmann, J., Exner, Y., & Domdey, S. (2024). Can Generative AI Create Superhuman Visual Marketing Content?

Highams, L., & Olszewska, J. I. (2023). Markov-Chain-Based Musical Creative Intelligent Agent Passing Successfully the Turing Test. *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*.

Hultén, B., Broweus, N., & van D. M. (2009). *Sensory Marketing*. Palgrave Macmillan.

IFPI. (2024). IFPI GMR 2024 State of the Industry.

Jones, P. W. (2021). Can Anybody Hear Me? The Power of Music in Advertising. *Kantar*. Retrieved March 20, 2026, from https://www.kantar.com/inspiration/advertising-media/can-anybody-hear-me-the-power-of-music-in-advertising

Keller, K. L. (1984). Advertising Retrieval Cues on Brand Evaluations. *Journal of Consumer Research*.

Kerr, A. H., & Das, N. (2014). Use of Music in Marketing Strategy. *Atlantic Marketing Journal*.

Kramer, K. (2024). "Verknallt in einen Talahon" klettert auf Platz 48 der Charts. *faz.de*. Retrieved March 20, 2026, from https://www.faz.net/aktuell/feuilleton/debatten/ki-song-verknallt-in-einen-talahon-in-top-50-augenzwinkern-oder-diskriminierung-19928947.html

Lin, T.-F., & Chen, L.-B. (2024). Harmony and Algorithm: Exploring the Advancements and Impacts of AI-Generated Music. *IEEE Potentials*, 2–9. https://doi.org/10.1109/mpot.2024.3433888

McKay, B. (2023). TV Marketing's Reign Is Over; Now Social Media Has Taken Its Place. *Forbes*. Retrieved March 20, 2026, from https://www.forbes.com/councils/forbesagencycouncil/2023/12/01/tv-marketings-reign-is-over-now-social-media-has-taken-its-place/

Monem, D. H. N. A. (2021). The Effectiveness of Advertising Personalization. *Journal of Design Sciences and Applied Arts*, *2*(1), 335–344. https://doi.org/10.21608/jdsaa.2021.31121.1061

Neese, K. V. (2015). The Role of Music in the Enhancement of Marketing.

Nelson, S. (2024). Red Lobster Uses AI to Write 30 Songs About Cheddar Bay Biscuits. *Adweek*. Retrieved March 20, 2026, from https://www.adweek.com/creativity/red-lobster-ai-songs-cheddar-bay-biscuits

NoPriors-Podcast. (2024). No Priors Ep. 64 | With Suno CEO and Co-Founder Mikey Shulman.

Persado. (2019). JPMorgan Chase Announces Five-Year Deal with Persado for AI-Powered Marketing Capabilities. *Persado Press Release*. Retrieved March 20, 2026, from https://www.persado.com/press-releases/jpmorgan-chase-announces-five-year-deal-with-persado-for-ai-powered-marketing-capabilities/

Pham, M. (1998). Representativeness, Relevance, and the Use of Feelings in Decision Making. *Journal of Consumer Research*, *25*(2), 144–159.

Rogès, N. (2024). The Music Industry Is at War Against AI: Suno and Udio's Case. *SoundiizBlog*.

Rosenfeld, A. (1985). Music, the Beautiful Disturber; Whether It's Bach, Beatles, the "Boss", Blues or Ballads, Chances Are That Music Speaks to Your Emotions. *Psychology Today*.

Sandred, O. R., Laurson, M., & Kuuskankare, M. (2009). Revisiting the Illiac Suite - A Rule-Based Approach to Stochastic Processes. *Sibelius Academy*.

Schneider, P. (2021). Künstliche Intelligenz – die deutschen Verbraucher bleiben skeptisch. *YouGov.de*.

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments. *Marketing Science*, *36*(4), 500–522. https://doi.org/10.1287/mksc.2016.1023

Silva, P. C. D. (2003). *Experiments in Musical Intelligence*. Spectrumpress.

Soundful. (2024). The Role of AI Music Generation for Content Creators. Retrieved March 20, 2026, from https://soundful.com/ai-music-generation-content-creators/

Stability.ai. (2024). Introducing Stable Audio 2.0.

StableAudio.com. (2024). Stable Audio. Retrieved March 20, 2026, from https://stableaudio.com/

Suno.com. (2024). Suno. Retrieved March 20, 2026, from https://suno.com/

Tencer, D. (2024). After Raising $125m, AI Music Generator Suno Is Now Paying Its Most Popular Creators. *Music Business Worldwide*. Retrieved March 20, 2026, from https://www.musicbusinessworldwide.com/after-raising-125m-ai-music-generator-suno-is-now-paying-its-most-popular-creators/

Tom, G. (1990). Marketing with Music.

WPP. (2022). AKQA: Nike's Never Done Evolving. Retrieved March 20, 2026, from https://www.wpp.com/en/featured/work/2022/08/akqa-nikes-never-stop-evolving
Zenieris, R. (2023). Perception and Bias Towards AI Music.

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Predicting Stock Market Trends Using Convolutional Neural Networks: A Deep Learning Approach

Ege Özkul

Technical University of Munich

Abstract

Technical analysis aims to predict stock returns based on price and volume patterns and has seen growing adoption of machine learning methods. However, most approaches rely on hand-crafted features. This paper investigates whether deep learning applied to stock chart images can predict future returns without manual feature engineering. Building on Jiang et al. (2023), this study applies Convolutional Neural Networks (CNNs), a computer vision architecture, to predict stock returns from price charts and extends their approach by implementing Vision Transformers, specifically the Class-Attention in Image Transformer (CaiT). Stock prices, volumes, and moving averages are encoded into images, which are used to train models that classify future stock returns as either positive or negative. Results show that both CNN and CaiT models outperform traditional technical indicators such as momentum and reversal strategies when applied to US stocks. Moreover, combining the two models yields incremental predictive power. An investment strategy based on their joint predictions achieves higher returns and Sharpe ratios than either model alone.

Keywords: technical analysis; stock return prediction; deep learning; convolutional neural networks; vision transformers

1 Introduction

Technical analysis is a method used to forecast future asset price movements based on the examination of past market data, primarily price and volume (Brock et al., 1992, p. 1731). Rather than relying on economic fundamentals, technical analysis identifies patterns and trends in historical price behavior, assuming that market movements exhibit recurring and predictable structures. This approach can be applied qualitatively, through the interpretation of recurring price formations on charts, or quantitatively, using technical indicators such as moving averages, which smooth out price fluctuations over time to reveal underlying trends (Menkhoff & Taylor, 2007, p. 939). This reliance on historical price patterns and market trends directly contradicts the Efficient Market Hypothesis (EMH), particularly its weak form (Murray et al., 2024, p. 1). The weak form EMH suggests that price changes follow a random walk, meaning that past price movements do not provide any predictive power for future returns (Fama, 1970, p. 389). In contrast, technical analysis assumes that historical price behavior influences future price movements, which implies a degree of market inefficiency.

Technical analysis is widely used in the financial industry (Zhu & Zhou, 2009, p. 519). Most brokerages provide charting analysis, making technical tools more accessible to market participants (Han et al., 2021, p. 1). Menkhoff and Taylor (2007) highlight its widespread adoption among foreign exchange traders, while (Lo & Hasanhodzic, 2010) examine how professional practitioners implement technical strategies in real-world trading. However, despite its extensive use in the industry, academics have traditionally doubted the validity of technical analysis, often questioning its reliability in predicting asset prices. Some even compare it to astrology or dismiss it as "voodoo finance," emphasizing the skepticism surrounding its legitimacy (Lo et al., 2000, p. 1705). Nevertheless, several studies have provided strong empirical evidence supporting technical analysis such as (Brock et al., 1992; Jegadeesh & Titman, 1993; Lo et al., 2000) and (Zhu & Zhou, 2009).

More recently, machine learning models have been increasingly applied in technical analysis. One of the biggest challenges in utilizing these techniques is the need for feature engineering, as predictive signals must be manually designed before they can be effectively used (Jiang et al., 2023, p. 3193). To address this limitation, (Jiang et al., 2023), in their paper "(Re-)Imag(in)ing Price Trends" propose a novel approach that converts stock data into images and applies Convolutional Neural Networks (CNNs) to automatically learn predictive features without manual intervention. Their findings show that image-based predictive features significantly enhance return predictions compared to traditional trend signals. The CNN models identified previously unexplored price patterns that proved

I would like to thank Professor Dr. Christoph Kaserer and Minghui Chen for their valuable guidance and support throughout the preparation of this thesis.

DOI: https://doi.org/10.5282/jums/v11i1pp195-226

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

more effective in forecasting returns, ultimately leading to superior portfolio performance.

This paper replicates their study by implementing the proposed CNN models and extends it by integrating a state-of-the-art image recognition architecture: Vision Transformers (ViTs; (Dosovitskiy et al., 2020)). ViTs are built on the Transformer architecture, which is also used in OpenAI's ChatGPT (Raschka, 2024, p. 6). Unlike CNNs, which process images by detecting local patterns such as edges and textures, ViTs divide images into patches and use a self-attention mechanism to learn relationships between different regions of the image. This allows ViTs to capture long-range dependencies and global context within an image, making them particularly effective for tasks requiring a broader understanding of spatial relationships (Maurício et al., 2023, p. 13).

Jiang et al. (2023) argue that one of the reasons for representing time series as images is that it allows models to capture structural relationships that may be challenging to identify using conventional time-series approaches, much like how humans recognize trends more easily in visual graphs than in raw numerical data. By utilizing this structured format, an advanced pattern detection model such as CNN can process the entire visual representation, enabling it to uncover hidden dependencies and trends that might otherwise go unnoticed Jiang et al. (2023, p. 3194). Building on this logic, a model designed to capture broader contextual relationships across an entire image may extract even more meaningful insights. Given that CNNs specialize in detecting local features such as edges and textures, they excel at identifying small-scale visual patterns within stock price images. However, ViTs may further enhance pattern detection in stock charts. By analyzing the image holistically rather than focusing on localized features, ViTs could improve the model's ability to recognize overarching trends and structural formations in stock charts. This paper tests whether the image-based approach introduced by (Jiang et al., 2023) can be further enhanced using ViTs.

Following (Jiang et al., 2023), I converted US stock data from 1993 to 2019 into 5-day, 20-day, and 60-day images. Each image is in grayscale format, incorporating price data, trading volume, and moving averages. For the ViT models, I adapted the images to an RGB format, where each color channel represents either price data, volume, or moving averages. Jiang et al. (2023) trained models on three image types (5-day, 20-day, and 60-day) to predict whether stock returns would be positive or negative over the subsequent week, month, or quarter, resulting in nine models. In my implementation, I trained eight models, omitting the 60-day image model for weekly return predictions due to storage constraints. For the ViTs, I trained only three models: two predicting weekly returns using 5-day and 20-day images, and one predicting quarterly returns using 5-day images. Given the high computational cost of training ViTs, I limited their implementation to these three cases. I utilized the Class-Attention in Image Transformers (CaiT) architecture proposed by (Touvron et al., 2021), a specialized variant of ViT that allows for deeper network architectures. Unlike standard ViTs, CaiT first learns the relationships within the image and then leverages these learned relationships to make predictions (Touvron et al., 2021). Additionally, ViTs lack inductive bias which makes them difficult to train from scratch or with small datasets. To address this challenge, I applied two techniques proposed by (S. H. Lee et al., 2021).

Both CNN and CaiT models generate probability estimates indicating the likelihood of a stock having a positive return over the prediction period. Following (Jiang et al., 2023), stocks are sorted into decile portfolios based on these predictions, and the portfolios are rebalanced at the same frequency as the prediction horizon to ensure consistency in evaluation. The performance of CNN-based strategies is assessed against traditional trend following signals, including 2-12 momentum (MOM), monthly short-term reversal (STR), and weekly short-term reversal (WSTR). Strategy performance is measured using a High-Low (H-L) portfolio and Sharpe ratios. Furthermore, I compare the performance of CaiT-based strategies with their CNN-based counterparts.

After comparing the performance of CaiT and CNN strategies, I examined their potential for integration. Since each model may focus on different features within the same image, combining them could yield superior results. To test this hypothesis, I first calculated the simple average of both models' probability estimates and sorted stocks into deciles based on these averages. This averaging process introduces an element of agreement in the High (highest decile or decile 10) and Low (lowest decile or decile 1) deciles, ensuring that stocks assigned a high probability by both models are positioned in the High decile, while those with low probabilities fall into the Low decile. Analyzing these deciles along with the H-L portfolio provides insights into the combined behavior of the models.

Building on this consensus-based approach, I propose an investment strategy that leverages both models by selecting stocks where they agree most strongly. I evaluate the strategy by reporting annualized returns and Sharpe ratios, assessing its resilience to transaction costs, and analyzing its cumulative returns over the out-of-sample period. This study also investigates whether CNNs and ViTs can be effectively integrated to enhance stock return predictions. To the best of my knowledge, no prior research has explored the use of stock charts as inputs for ViTs to predict stock returns or attempted to combine ViTs and CNNs for improved return prediction performance.

The remainder of this thesis is structured as follows: Section 2 reviews the relevant literature on Technical Analysis and Computer Vision. Section 3 introduces the fundamental concepts of CNNs and ViTs, detailing the architectural design choices. Section 4 outlines the data sources and methodology employed in implementing both architectures. Section 5 presents the empirical results for CNN and CaiT models, including the effects of averaging their predictions. Section 6 proposes the investment strategy derived from these models. Section 7 discusses the limitations of the study and provides a critical evaluation of the findings. Finally, Section 8 concludes with key insights.

2 Literature Review

This paper explores the intersection of technical analysis, Convolutional Neural Networks, and Vision Transformers. It investigates how CNN-based image analysis can enhance technical analysis and then applies ViTs to assess whether stock prediction can be further improved using an alternative image recognition algorithm. Accordingly, this section reviews the relevant literature from two perspectives: first, technical analysis, and second, computer vision.

Technical analysis has a long history, with its origins traced back to Charles Dow in the late 1800s in the United States Brock et al. (1992, p. 1731). Dow developed what became known as Dow Theory, which aimed to predict market trends by identifying upward or downward price movements Zhu and Zhou (2009, p. 521). However, early empirical studies cast doubt on the effectiveness of technical analysis. Cowles (1933) conducted one of the first major empirical tests, analyzing the forecasts of William Peter Hamilton, a leading Dow Theory advocate. Cowles (1933) found that Hamilton's predictions between 1904 and 1929 underperformed a simple buy-and-hold strategy, suggesting that technical forecasting lacked predictive power. Later studies reinforced this skepticism—(Fama & Blume, 1966) tested 24 filter rules and found that they did not outperform passive investing, while (Jensen & Benington, 1970) examined Levy (1967)'s Relative Strength Index (RSI) and found that it did not yield superior returns once transaction costs were accounted for. The Efficient Market Hypothesis, introduced by (Fama, 1970), further challenged technical analysis by asserting that past prices contain no useful information for predicting future movements—a direct contradiction to the principles of technical analysis. As a result, early studies were generally skeptical of technical analysis, casting doubt on its ability to generate consistent excess returns.

Despite early skepticism, more recent research has provided empirical evidence supporting the effectiveness of technical analysis in financial markets. Lo et al. (2000) developed a quantitative framework to systematically detect technical patterns and rigorously tested them, finding that technical analysis can offer incremental information beyond random price movements, suggesting that market inefficiencies may exist in the short term. Jegadeesh and Titman (2001) found that momentum strategies remain highly profitable and are not a product of data snooping. Their results indicate that systematic mispricing, driven by behavioral biases, can create exploitable trading opportunities. Similarly, (Neely et al., 2014) examined the relationship between technical indicators and the equity risk premium, finding that technical analysis and macroeconomic variables capture different aspects of market dynamics. Their research shows that combining the two yields superior out-of-sample forecasts, challenging the traditional view that only fundamental data is valuable. More recently, (Murray et al., 2024) tested the weak-form Efficient Market Hypothesis using machine learning (ML) models trained on historical price data. Their findings suggest that ML-based technical analysis strategies generate significant excess returns, which cannot be fully explained by risk factors, further supporting the idea that technical patterns may contain predictive power. Collectively, these studies indicate that there is strong empirical evidence supporting the effectiveness of technical analysis.

Moving averages (MAs) serve as the foundation of trend-following strategies, which are widely used in financial markets Han et al. (2021, p. 1). One of the first major studies on MAs, conducted by (Brock et al., 1992), analyzed 90 years of Dow Jones Industrial Average data and found that MAs and Trading Range Breakouts exhibited statistically significant predictive power. However, (Sullivan et al., 1999) revisited this finding, testing whether MAs truly predict stock returns or if their past performance was due to data-snooping bias. Their results showed that while MAs appeared profitable in historical data, they lost predictive power in out-of-sample tests. Despite this skepticism, later studies still found empirical support for MAs. Zhu and Zhou (2009) examined MAs from an asset allocation perspective and concluded that they help investors make better allocation decisions, especially in uncertain market conditions. Han et al. (2016) introduced the Trend Factor, a model that integrates MAs across different time horizons to capture persistent price trends, making them particularly effective during market uncertainty. More recently, (Detzel et al., 2021) examined MAs in markets where fundamental valuation is difficult, such as Bitcoin and speculative stocks, and found that MA-based strategies generate significant excess returns, particularly in cryptocurrencies. These studies suggest that MAs remain a valuable tool for traders and investors.

In addition to traditional technical analysis methods, recent research has increasingly incorporated advanced machine learning techniques such as Support Vector Machines (SVMs) and Genetic Algorithms (GAs) to improve predictive accuracy in financial markets. SVMs, first introduced by (Noble, 2006), classify data by finding an optimal hyperplane, making them effective in financial forecasting. Cao and Tay (2003) demonstrated that SVMs with adaptive parameters outperform Backpropagation Neural Networks when predicting futures contracts from the Chicago Mercantile Market. M. Lee (2009) further refined SVM-based forecasting by integrating a hybrid feature selection method with financial data from NASDAQ, 20 futures contracts, and 9 spot indexes, achieving superior accuracy and generalization. Additionally, (Kercheval & Zhang, 2015) leveraged high-frequency limit order book data from NASDAQ to build multi-class SVM models, successfully predicting short-term price movements and generating profitable trading signals in real-time. Beyond SVMs, Genetic Algorithms, first developed by (Holland, 1975), offer an evolutionary approach to financial forecasting by optimizing trading strategies through natural selection principles. Allen and Karjalainen (1999) pioneered the use of GAs in technical analysis, applying them to develop trading rules for the S&P 500, though results indicated they did not consistently outperform a buy-and-hold strategy after transaction costs. More recent studies have improved GA-based approaches: (Brown et al.,

2013) introduced the Dynamic-Radius Species-Conserving Genetic Algorithm to select stocks from the Dow Jones Index, significantly outperforming the index in returns. Y. Hu et al. (2015) combined GAs with the eXtended Classifier System in their eTrend model, which adapts trading rules dynamically, outperforming buy-and-hold strategies with superior risk-adjusted returns.

Neural networks have become a widely used approach in financial forecasting due to their ability to model complex, nonlinear relationships in stock price data. Tsang et al. (2007) developed NN5, a Backpropagation Neural Network to predict HSBC stock price movements in the Hong Kong market, achieving over 70% accuracy, though profitability was limited by transaction costs. Similarly, (Guresen et al., 2011) compared various Artificial Neural Network architectures for stock market index prediction, and found that the Multi-Layer Perceptron produced the most accurate and stable forecasts for NASDAQ index movements. Further advancing the field, (Nelson et al., 2017) applied Long Short-Term Memory (LSTM) networks to predict stock price movements on the Brazilian stock exchange using historical price data and technical indicators, demonstrating LSTM's superiority over traditional machine learning models, with an accuracy of up to 55.9% in short-term price direction prediction. These studies highlight the growing role of neural networks in technical analysis, offering improved predictive power and adaptability compared to conventional methods.

Recent research has explored the application of CNNs in financial market prediction, with some studies integrating them with other models while others use CNNs independently. Kim and Kim (2019) developed a hybrid LSTM-CNN model, showing that combining numerical and image-based stock data outperforms standalone CNNs or LSTMs. Similarly, (Di Persio & Honchar, 2016) introduced a Wavelet-CNN model, demonstrating that wavelet preprocessing enhances feature extraction and that ensemble learning further improves the accuracy of stock market price prediction. J. Lee et al. (2019) applied CNNs within a Deep Q-Network, demonstrating that U.S.-trained CNNs generalize well to global markets, while (Hoseinzade & Haratizadeh, 2019) found that multi-market CNNs outperform single-market models for stock index movement prediction. In contrast, other studies relied solely on CNNs, using either raw numerical data or transformed financial images. Gunduz et al. (2017) enhanced CNN accuracy by structuring input features according to their correlation, and (G. Hu et al., 2018) used Convolutional AutoEncoders to extract deep features from candlestick charts, demonstrating better stock clustering for portfolio optimization. Similarly, (Cohen et al., 2020) used CNN-based image classification to replicate human technical analysis, achieving 95% accuracy in detecting buy signals, while (Chen et al., 2016) showed that CNNs outperformed traditional market models in time-series forecasting. However, (Jiang et al., 2023) provided the most comprehensive demonstration of CNNs' predictive power, transforming stock price trends into images and proving that CNNs outperform traditional momentum and reversal strategies, generalize across time scales, and successfully transfer from U.S. to global markets.

Convolutional Neural Networks have played a crucial role in the advancement of image recognition. Introduced by (Lecun & Bengio, 1995), CNNs leverage spatial hierarchies and local connectivity to effectively process image data. Over the years, several architectures have significantly contributed to the field. AlexNet, proposed by (Krizhevsky et al., 2012), was a groundbreaking deep CNN that won the ImageNet competition in 2012, demonstrating the power of deep learning in large-scale image classification. It featured multiple convolutional layers, max-pooling, and the use of ReLU activations to speed up training and improve accuracy. GoogleNet, developed by (Szegedy et al., 2014), introduced the Inception module, which optimized computational efficiency while increasing network depth and width, allowing for multi-scale feature extraction. ResNet, proposed by (He et al., 2016a), addressed the vanishing gradient problem in deep networks by introducing residual connections, enabling the training of networks with over 100 layers while maintaining high accuracy. These architectures have set benchmarks for deep learning-based image analysis and continue to inspire modern developments in computer vision.

More recently, Vision Transformers have emerged as a powerful alternative to CNNs for image recognition. Inspired by the Transformer architecture originally proposed by (Vaswani et al., 2017) for natural language processing, ViTs were introduced by (Dosovitskiy et al., 2020) as a model that treats images as sequences of patches, similar to words in a text sequence. Unlike CNNs, which rely on convolutional layers to extract local features, ViTs use self-attention mechanisms to model global dependencies within an image, allowing them to capture long-range interactions Maurício et al. (2023, p. 13). However, early ViTs required large-scale datasets for effective training, which led to the development of optimized versions such as Data-efficient Image Transformers (DeiT) by (Touvron et al., 2020), which introduced a knowledge distillation approach to make ViTs more accessible with limited data. Other variations like Pooling-based Vision Transformers (Heo et al., 2021) incorporated spatial dimension reduction inspired by CNNs to improve efficiency, and Class-Attention in Image Transformers by (Touvron et al., 2021) introduced specialized attention mechanisms to enhance classification accuracy. These advancements have solidified ViTs as a competitive alternative to traditional CNNs in image analysis tasks, pushing the boundaries of deep learning in computer vision.

To address the limitations of ViTs on small datasets, various strategies have been proposed to enhance their performance without extensive pre-training on large-scale datasets. One such approach by (S. H. Lee et al., 2021) introduces Shifted Patch Tokenization and Locality Self-Attention to improve locality inductive bias, making ViTs more effective for small-scale datasets. SPT increases the receptive field of tokens, allowing ViTs to capture more spatial relationships between neighboring pixels, while LSA

refines the attention mechanism to focus more on relevant local features rather than distributing attention uniformly. Similarly, (Shao & Bi, 2022) propose a hybrid transformer model that integrates Convolutional Parameter Sharing Multi-Head Attention and Local Feed-Forward Networks, effectively combining the strengths of CNNs and transformers. Their method incorporates convolutional layers within the transformer structure, improving local feature extraction while maintaining the long-range dependencies of self-attention. Additionally, (Gani et al., 2022) present a self-supervised weight initialization approach, leveraging low-resolution feature prediction to pre-train ViTs on small datasets without relying on external large-scale datasets. These advancements contribute to making ViTs more practical for small-scale vision tasks, reducing dependency on massive datasets while improving performance and efficiency.

3 Theoretical Background

In this section, I will discuss the theoretical background of CNNs and ViTs. This section has two main purposes: first, to introduce these algorithms, explain their fundamental mechanics, and highlight the key ideas behind them; second, to illustrate how they apply to stock return prediction and the rationale behind my design choices.

I will begin with CNNs, explaining their general structure and core concepts. Next, I will discuss Transformers, the foundation of Vision Transformers, and explain the self-attention mechanism from a natural language processing perspective. Then, I will describe how this mechanism is adapted to images by outlining the standard ViT architecture and its potential to improve pattern recognition in stock charts. Following this, I will introduce the specific CaiT architecture and justify its selection for this task. Finally, the chapter will conclude with an explanation of two techniques that enable ViTs to be trained from scratch on stock image datasets, making their application feasible for this study.

3.1 CNN Background

Convolutional Neural Networks are a specialized class of neural networks designed to process data with grid-like structures, such as images Goodfellow et al. (2016, p. 330). In a standard CNN layer, there are three key stages: the convolution stage, involves applying learned filters to small sections of the input data; the activation stage, which applies a non-linear function to introduce non-linearity; and the pooling stage, groups nearby values together, simplifying the feature maps and helping the network ignore small changes in the input Goodfellow et al. (2016, p. 339). Furthermore, CNNs leverage two fundamental principles: parameter sharing, which allows the same filter to be used across different regions, reducing both the number of parameters and the amount of computation required; and translation equivariance, meaning that shifts in the input lead to corresponding shifts in the output, thereby maintaining the spatial consistency of the detected features Jiang et al. (2023, p. 3240).

In the convolution operation, a small matrix of learned parameters, known as a kernel or filter, slides over the input data to compute weighted sums at every position, thereby detecting local patterns such as lines or edges Goodfellow et al. (2016, p. 335). This process is used to transform raw input into a more informative representation that highlights important features while preserving spatial relationships Jiang et al. (2023, p. 3238). Figure 1 demonstrates this process: each element of the kernel (labeled w, x, y, z) is multiplied by corresponding input values (labeled a, b, c, etc.) within a specific patch, and these products are summed to yield a single output value. As the kernel moves across the entire input, it produces an output map that indicates where the feature appears and with what intensity. Importantly, the weights of these filters are learned during training, enabling the network to adapt to the specific patterns present in the data, and in practice, many different filters are applied—not just one—to capture a wide variety of features essential for effective analysis (Goodfellow et al., 2016, pp. 333–336).

Following the convolution operation, an activation function is applied element-wise to the resulting feature maps to introduce non-linearity Jiang et al. (2023, p. 3240). This non-linearity is crucial because, without it, the entire network would collapse into a simple linear model, irrespective of its depth Goodfellow et al. (2016, p. 172). By incorporating non-linearity, neural networks gain the ability to approximate any function, enabling them to capture intricate relationships between inputs and outputs when given sufficient layers. This transformation allows networks to convert linear input data into highly complex, non-linear outputs, making them effective for modeling complex structures that would be unattainable with a purely linear approach Goodfellow et al. (2016, p. 198). One of the most common activation functions is the Rectified Linear Unit (ReLU; Goodfellow et al. (2016, p. 193)), which is defined as:

$$f(x) = \max(0, x) \tag{1}$$

Following the convolution and activation steps, the pooling stage aggregates information from local regions of the activation maps to simplify the overall representation. In particular, max pooling partitions the feature maps into small, non-overlapping regions and selects the highest value from each region Goodfellow et al. (2016, p. 339). This operation reduces the spatial dimensions of the data, lowering computational complexity for subsequent layers, while also emphasizing the most prominent features. By doing so, max pooling makes the network more robust to slight shifts or distortions in the input, ensuring that key patterns are retained even if their exact locations change slightly Jiang et al. (2023, p. 3241).

As mentioned earlier, the convolution operation involves sliding a filter over the input data, and the same filter is used

[1] taken from Goodfellow et al. (2016, p. 334)

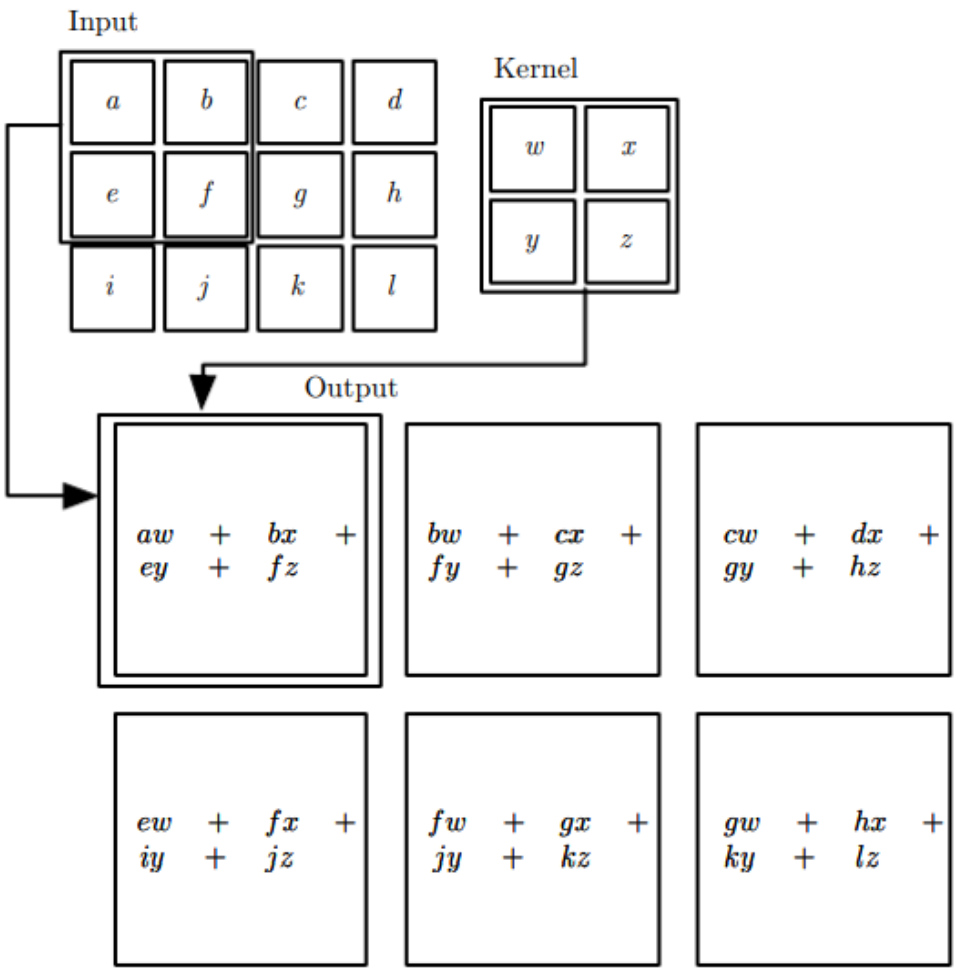

Figure 1: Convolution Operation Example[1]

Description. This figure illustrates the convolution operation in a neural network, where a 2D input matrix is processed using a small kernel (filter) to produce an output feature map. The input consists of a 4×3 matrix, while the kernel is a 2×2 matrix. The kernel slides over the input matrix, computing element-wise multiplications followed by summations to generate output values. Each resulting value represents a weighted sum of overlapping regions between the input and the kernel.

at every location. This technique is known as parameter sharing. The reuse of filter weights means that the network does not have to learn separate parameters for every region of the input, which significantly reduces the total number of parameters Goodfellow et al. (2016, p. 335). Not only does this make the model more memory- and computation-efficient, but it also reinforces translation equivariance—meaning that when an input is shifted, the output shifts in the same way Goodfellow et al. (2016, p. 338). This property ensures that if a feature is detected in one part of the image, it will still be recognized even if it appears in a different location, which is highly desirable for tasks like image recognition. Translation equivariance contributes to the robustness and generalization of the network by enabling consistent feature detection regardless of the location Jiang et al. (2023, p. 3240).

Now that we have examined key concepts of CNNs, an important consideration is the distinction between local and global features in images. Through the convolution operation, CNNs excel at detecting local features by analyzing small parts of an image, and due to translation equivariance, they can recognize these features—like eyes—regardless of their position Goodfellow et al. (2016, p. 342). However, capturing global features is more challenging. In a typical CNN, different sections of an image are not directly connected; instead, they are gradually linked through successive convolution layers Peng et al. (2023, p. 9455). Each convolution reduces the size of the output relative to its input, and although subsequent convolutions cover larger areas of the image, this information is aggregated indirectly Goodfellow et al. (2016, p. 337). Moreover, the application of max pooling further diminishes spatial information, as it selects only the maximum value from a given region, thereby losing precise details about where that value originated Goodfellow et al. (2016, p. 347). For traditional image recognition tasks,

such as face recognition, this trade-off in spatial precision is often acceptable. Instead of recognizing an entire face as a structured entity, CNNs can identify and classify faces based on the presence of key local features, such as eyes, nose, and mouth, and their approximate locations. The network does not require an exact representation of a face to correctly classify an image as containing a face Goodfellow et al. (2016, p. 342). However, in the context of stock charts, where the location of a feature also represents a specific time, preserving this spatial (and temporal) information becomes crucial. For example, from a technical analysis perspective, it is really important to detect the intersection of the moving average line and close price, but also where this intersection happened is also important. Recognizing different parts of a chart, such as the beginning and end of a period, is crucial, as they may convey distinct information. This understanding led me to implement an alternative image recognition model to better capture global patterns.

3.2 Transformer Background

Vision Transformers are image analysis models built on the Transformer architecture, sharing the fundamental mechanism of self-attention. While Transformers were originally developed for natural language processing (NLP), ViTs adapt this architecture to process images as input. Beyond this key difference, the core principles remain largely the same. Therefore, in this section, I will first introduce Transformers and the self-attention mechanism from a natural language perspective. This foundational understanding will serve as a basis for the following section, where I will explain Vision Transformers by building upon the concepts introduced here.

Transformers represent a major breakthrough in deep learning, having fundamentally reshaped the field of NLP since their introduction by (Vaswani et al., 2017). Through an attention mechanism, the Transformer architecture combines the computational parallelism of Convolutional Neural Networks with the ability to model long-range dependencies and process variable-length sequences, which provides an advantage for the long texts—a strength traditionally associated with Recurrent Neural Networks (RNNs). This innovative approach not only boosts efficiency during training but also leads to superior performance across various tasks Kamath et al. (2022, p. 19).

Before a Transformer model can process text, it must first convert raw language into a numerical format that neural networks can understand. Text cannot be directly processed mathematically for training deep learning models because it consists of categorical data. To bridge this gap, words and symbols are represented as continuous-valued vectors through a process known as embedding Raschka (2024, p. 25).

Before embedding occurs, an essential preprocessing step called tokenization is performed. Tokenization breaks the input text into discrete units, known as tokens, which can be individual words, subwords, or special characters, including punctuation Raschka (2024, p. 27). These tokens are then converted into numerical representations, typically integers, which serve as a unique ID and will be the input to the embedding layer Raschka (2024, p. 33).

At its core, an embedding maps discrete tokens into a continuous vector space. In the embedding layer, each token is associated with a high-dimensional vector, drawn from a structured matrix known as the embedding matrix Kamath et al. (2022, p. 20). This matrix consists of a set number of vectors, each corresponding to a unique token in the model's vocabulary Raschka (2024, p. 34). The dimensionality of these vectors is predefined and chosen based on the model's design requirements. OpenAI's GPT-3, for example, uses embeddings with 12,288 dimensions Raschka (2024, p. 28). Higher-dimensional embeddings allow for more nuanced distinctions between words, capturing deeper semantic relationships Courant et al. (2023, p. 199). When a set of tokens enters the embedding layer, the model uses their unique IDs to locate their corresponding vectors within the embedding matrix, returning the appropriate embeddings Raschka (2024, p. 56). Initially, the values in this matrix, known as weights, are assigned randomly. As training progresses, these values are iteratively refined through backpropagation, allowing the model to adjust embeddings based on conceptual relationships found in the data Raschka (2024, p. 54). This optimization process ensures that words frequently appearing in similar conceptual categories develop embeddings that move closer together in the vector space. For example, as training continues, vectors for "king" and "queen" become more similar, as do those for "dog" and "cat," reflecting their semantic connections Raschka (2024, p. 28).

One important characteristic of Transformers is that they are permutation invariant, meaning they do not inherently recognize word order Lin et al. (2022, p. 122). For example, the phrases "Fox jumps over dog" and "Dog jumps over fox" would appear identical to a Transformer Raschka (2024, p. 57). To address this, positional embeddings are added to token embeddings, encoding the position of each token in a sequence, such as a sentence, to provide a sense of order Lin et al. (2022, p. 122).

By converting discrete tokens into embedding vectors, Transformers can process text in a mathematically meaningful way, forming the foundation for all subsequent learning. Since these embeddings are trainable, they evolve over time, allowing the model to continuously improve its language understanding as it learns from more data Raschka (2024, p. 68).

The self-attention mechanism is the most important component of a Transformer. It enables the model to focus on long sequences of text and capture contextual relationships between words effectively. Unlike traditional sequence-based models, which process tokens sequentially, self-attention allows each token to analyze its relationship with all other tokens in the sequence simultaneously Raschka (2024, p. 70).

To achieve this, the self-attention mechanism relies on three weight matrices: query mapping, key mapping, and value mapping. These matrices contain learnable parameters that are updated during training Courant et al. (2023, p. 195). For each token in a sequence, its embedding vector

is multiplied by these matrices, producing the query, key, and value representations. The query vector of a token is then compared with the key vectors of all tokens, including itself, to compute similarity scores using dot products Courant et al. (2023, p. 195). A high dot product value indicates a strong relationship, whereas a lower value suggests a weaker connection Raschka (2024, p. 73).

These similarity scores, often referred to as attention scores, determine how much focus each token should place on others in the sequence. The attention scores are then normalized using a softmax function, converting them into attention weights. These weights dictate the influence each token's value vector has on the final representation of the token being processed (Raschka, 2024, pp. 72–74).

To illustrate this process, consider the sentence: "Apple is a fruit." This sentence consists of four tokens: "Apple", "is", "a", "fruit". To calculate the final representation of "Apple" and determine how much it is influenced by other words, we first compute its attention scores by taking the dot product of the query vector of "Apple" with the key vectors of all tokens in the sequence, including itself. This results in four distinct attention scores for "Apple". These attention scores are then converted into attention weights using the softmax function Raschka (2024, p. 74).

Once the attention weights are obtained, they are used to compute the final representation of "Apple". Each attention weight is multiplied by the value vector of the corresponding token: the attention weight between "Apple" and "fruit" is multiplied by the value vector of "fruit", the attention weight between "Apple" and "is" is multiplied by the value vector of "is", and so on. The weighted value vectors are then summed, forming the final representation of "Apple", also referred as context vector, which represents how it interacts with the entire sequence. This process is performed simultaneously for every token in the sentence, allowing each word to build a context-aware representation Raschka (2024, p. 7).

By leveraging self-attention, Transformers efficiently capture dependencies between words, regardless of their distance in the sequence. This ability to dynamically adjust attention weights makes them highly effective in understanding language and processing complex sequences Courant et al. (2023, p. 195).

While self-attention allows Transformers to capture relationships between tokens, multi-head self-attention extends this capability by enabling the model to learn multiple types of relationships simultaneously. Instead of computing a single self-attention, multi-head self-attention applies several self-attention mechanisms in parallel, each with its own set of query, key, and value matrices. These independent attention heads allow the model to focus on different aspects of the input sequence Kamath et al. (2022, p. 24). After processing, the outputs of all attention heads are concatenated, ensuring that the model retains diverse contextual information. This mechanism enhances the Transformer's ability to understand complex language structures, improving its performance across various natural language processing tasks Raschka (2024, p. 101).

Multi-head self-attention is a crucial concept in the Transformer architecture, enabling the model to focus on different aspects of a sequence simultaneously. After multi-head self-attention, the output passes through layer normalization, a technique that stabilizes training by normalizing activations across the layer. This helps control the variance of inputs, ensuring that the network learns more efficiently and avoids issues like exploding or vanishing gradients Raschka (2024, p. 120).

Following layer normalization, the transformed embeddings are processed by a fully connected feedforward neural network (FNN). Unlike self-attention, which focuses on interactions between tokens, the FNN refines each token's representation independently Lin et al. (2022, p. 113). It consists of two linear layers with a non-linear activation function, allowing the model to capture complex features and improve its expressiveness. The FNN ensures that the token representations remain useful for deeper layers while maintaining the same input and output dimensions for seamless stacking of other layers Raschka (2024, p. 130).

Finally, residual connections are applied to both the self-attention and FNN outputs. These connections help prevent information loss and enable better gradient flow, making it easier to train deep architectures. By adding the original input back to the transformed output, residual connections ensure that important features are preserved while still benefiting from the transformations introduced by self-attention and the FNN (Raschka, 2024, pp. 132–133).

Together, these components—multi-head self-attention, layer normalization, the feedforward network, and residual connections—work together to form a Transformer block. Multiple Transformer blocks can be stacked on top of each other Raschka (2024, p. 140).

3.3 Vision Transformer Background

Now that I have explained the key mechanisms behind Transformers, I will introduce the fundamental aspects of Vision Transformers. Vision Transformers share a similar architecture to standard Transformers but include modifications tailored for image processing. ViTs were first proposed by (Dosovitskiy et al., 2020) that introduced adaptations necessary for handling visual data. In this section, I will outline the key improvements made for processing images by explaining the basic structure of ViT as proposed in (Dosovitskiy et al., 2020). While the architecture I implemented differs slightly from the original design, I will discuss my modifications and design choices in the next chapter.

In the previous chapter, I explained how Transformers for NLP process words by converting them into tokens, where each token is assigned a unique ID. The model then uses these IDs to retrieve corresponding embedding vectors from an embedding matrix, which consists of a fixed number of vectors, each representing a word in the training data. Since words

[2] taken from Dosovitskiy et al. (2020, p. 3)

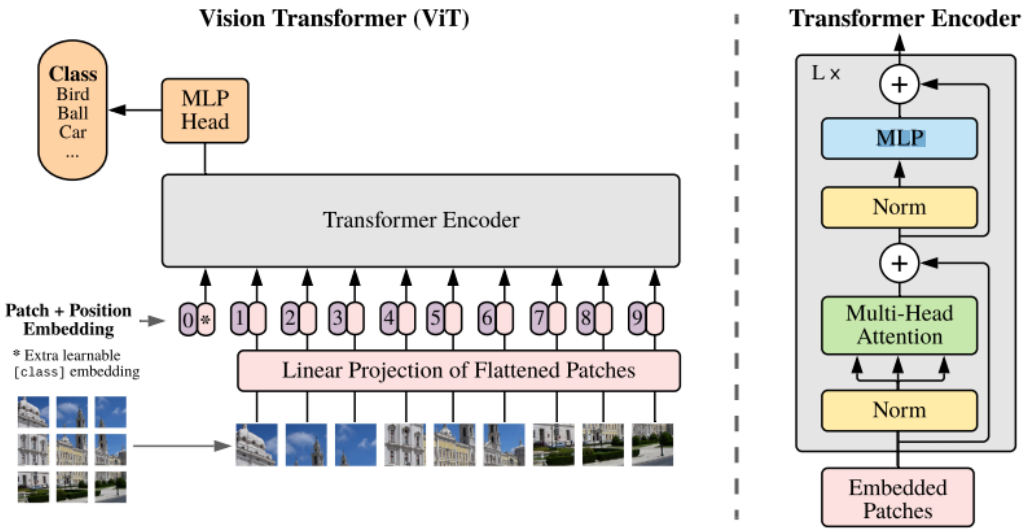

Figure 2: Vision Transformer Architecture[2]

Description. The figure illustrates the ViT pipeline on the left and the Transformer Encoder on the right. In the left part, an input image is divided into non-overlapping patches, flattened, and linearly projected into embeddings with positional encodings and a learnable class token. These embeddings are processed by the Transformer Encoder, and the class token output is passed to an Multi-Layer Perceptron (MLP) Head for classification. The right part shows the Transformer Encoder, where each layer follows a structured sequence: Layer Normalization (Norm) is applied first, followed by Multi-Head Self-Attention and a residual connection. Another Layer Normalization is then applied before passing through a MLP block, which is followed by a second residual connection.

are categorical, they always map to the same column in the embedding matrix, ensuring that a word like "eye" receives the same embedding regardless of the sentence it appears in. However, this approach does not directly apply to images, as there are countless pixel combinations that could resemble an eye. Unlike words, which have a predefined vocabulary, images do not have a fixed set of possible representations that can be directly mapped to an embedding matrix. Instead, Vision Transformers introduce patch embedding, where an image is divided into smaller patches rather than being processed as a whole Courant et al. (2023, p. 203). Figure 2 shows the basic ViT architecture that was proposed by (Dosovitskiy et al., 2020) and how the images are divided into patches. First the image is divided into nine patches. Each patch consists of raw pixel values that are first flattened into a one-dimensional vector. Instead of using an ID to retrieve an embedding, these flattened pixel vectors are mapped to a fixed-dimensional space through a trainable linear projection Dosovitskiy et al. (2020, p. 3). This projection functions similarly to an embedding matrix, ensuring that the transformed pixel values are in the appropriate format for the self-attention mechanism. During training, the model optimizes this projection, learning how to represent pixel information in a way that enhances its ability to process relationships between patches effectively Dosovitskiy et al. (2020, p. 4).

After generating patch embeddings, positional embeddings are added to retain spatial information, similar to how word positions are incorporated in traditional Transformers Dosovitskiy et al. (2020, p. 3). Just as positional embeddings help indicate word order in a sentence, they ensure that the model understands the relative arrangement of patches within the image. Additionally, a learnable class embedding, known as class token, is prepended to the sequence of patch embeddings Dosovitskiy et al. (2020, p. 3). Since this class token will also participate in the self-attention mechanism later, it will learn relationships between different patches, gradually aggregating information from the entire image. This token is designed to serve as a comprehensive representation of the entire image, making it particularly useful for classification tasks Courant et al. (2023, p. 203).

The prepared embeddings, consisting of the class token, patch embeddings, and positional embeddings, are then passed into a Transformer encoder Dosovitskiy et al. (2020, p. 3). The term "encoder" originates from the original Transformer architecture, which was designed for sequence-to-sequence tasks, such as language translation Vaswani et al. (2017, p. 3). Within the encoder, self-attention, specifically multi-head self-attention, plays a crucial role in capturing dependencies between different patches. Instead of treating patches independently, self-attention enables the model to analyze how different regions relate to one another along with the class token, effectively learning a global understanding of the image Touvron et al. (2020, p. 3). Just as self-attention helps capture context in sentences or paragraphs, it enables ViTs to extract meaningful relationships between distant patches in an image (Li et al. 2023, p. 1; Touvron et al. 2020, p. 3). This could be particularly beneficial for stock chart analysis, where the relationships

between different parts of the chart may contribute to accurate classification.

As seen in Figure 2, after self-attention learns relationships between patches, a Multilayer Perceptron refines the extracted information. Unlike traditional Transformer blocks, where layer normalization is applied after attention and feedforward layers, ViTs apply layer normalization before both multi-head self-attention and MLP layers. Additionally, residual connections are included to ensure stable gradient flow and efficient training Dosovitskiy et al. (2020, p. 3). Layer normalization, multi-head self-attention, MLP, and residual connections form the Transformer encoder in ViT. Dosovitskiy et al. proposed stacking 12 Transformer encoder layers as a baseline. Throughout these layers, the class token continuously updates by integrating information from important patches, ultimately forming a global representation of the entire image. After passing through all Transformer encoder layers, the final class token is sent to the MLP head, which generates the classification prediction Courant et al. (2023, p. 203).

To conclude, in ViTs, an image is divided into patches, and the pixel values of each patch are flattened. Through a trainable linear projection, the model learns how to represent these pixels as patch embeddings. The positions of the patches are also encoded and, along with their embeddings, are fed into the multi-head self-attention mechanism. Each self-attention head captures different relationships between the patches. For example, in the sentence 'I ate pizza, and I like it,' the self-attention mechanism learns that the word 'it' refers to 'pizza.' Similarly, in a stock chart image, different parts of the image learn how they relate to one another and whether those relationships influence stock returns. This ability to model dependencies across different regions of the image allows ViTs to potentially uncover patterns that CNNs might overlook.

3.4 Class-Attention in Image Transformer

In the previous chapter, I introduced the standard Vision Transformer architecture, which I will refer to as the base ViT. However, the architecture I implemented is Class-Attention in Image Transformers (CaiT), proposed by (Touvron et al., 2021). CaiT builds upon the base ViT with two key improvements: LayerScale and Class-Attention. These modifications address the limitations of base ViT, enabling deeper network training and improving classification efficiency.

Dosovitskiy et al. (2020) and (Touvron et al., 2020) have shown that increasing the number of layers does not necessarily enhance performance due to optimization challenges. CaiT overcomes this with LayerScale, a learnable diagonal matrix applied to residual connections Touvron et al. (2021, p. 4). By starting with small initial values and gradually learning the appropriate scaling, LayerScale stabilizes training, making it possible to train much deeper Transformers without convergence issues Touvron et al. (2021, p. 2).

The second improvement, Class-Attention, refines how the class token interacts with image patches. In base ViT, the class token attends to all patches at every layer, forcing the model to simultaneously learn both inter-patch relationships and classification features Touvron et al. (2021, p. 5). CaiT restructures this by first allowing self-attention layers to process only patch embeddings, learning spatial relationships without interference. The class token is introduced later in a dedicated class-attention stage, where it aggregates information from the learned patch representations Touvron et al. (2021, p. 5–6). This separation improves classification by ensuring that the class token captures global rather than prematurely biased local information (Touvron et al., 2021, pp. 20–21).

I chose CaiT for stock chart classification because its architectural modifications align with my assumptions about the nature of stock charts. Specifically, I assume that stock charts contain abstract patterns that hold predictive power over returns. In the base ViT, the class token is present from the beginning, aggregating information from different parts of the image. While this approach works well for natural images—where distinct visual features, such as a cat's eyes and ears, are immediately recognizable and crucial for classification—stock charts lack such obvious key elements. Instead, the patterns that influence returns are more subtle and complex. By postponing the introduction of the class token, CaiT allows the model to first focus on learning these abstract patterns within the image before associating them with the final prediction. Once the model has captured these patterns, the class-attention stage may then predict the return direction based on the fully learned representations. Given these assumptions, CaiT is a suitable choice. Additionally, in the appendix, I provide results from experiments with other architectures to compare their effectiveness.

3.5 Lack of Inductive Bias in ViT

Inductive bias refers to the set of assumptions a learning algorithm makes to generalize from training data to unseen data. Since a model cannot encounter all possible examples during training, it relies on educated guesses to make predictions. Inductive bias helps guide the model toward patterns or rules that are likely to work well, rather than treating all possible explanations equally. Without inductive bias, a model would be no better than random guessing when encountering new data Haussler (1988, p. 178).

CNNs have strong inductive biases, such as locality, where small filters focus on capturing local patterns, and translation equivariance, ensuring features are recognized regardless of their position in the image Battaglia et al. (2018, p. 6). However, these inductive biases are weaker in Vision Transformers compared to CNNs. In ViTs, self-attention layers are global, meaning spatially close patches do not inherently have stronger relationships—any two patches can be related, regardless of their position (Dosovitskiy et al., 2020, p. 4). Only MLP layers provide some degree of locality since they process each token independently. Additionally, MLP layers

[3] Modified, taken from S. H. Lee et al. (2021, p. 3).

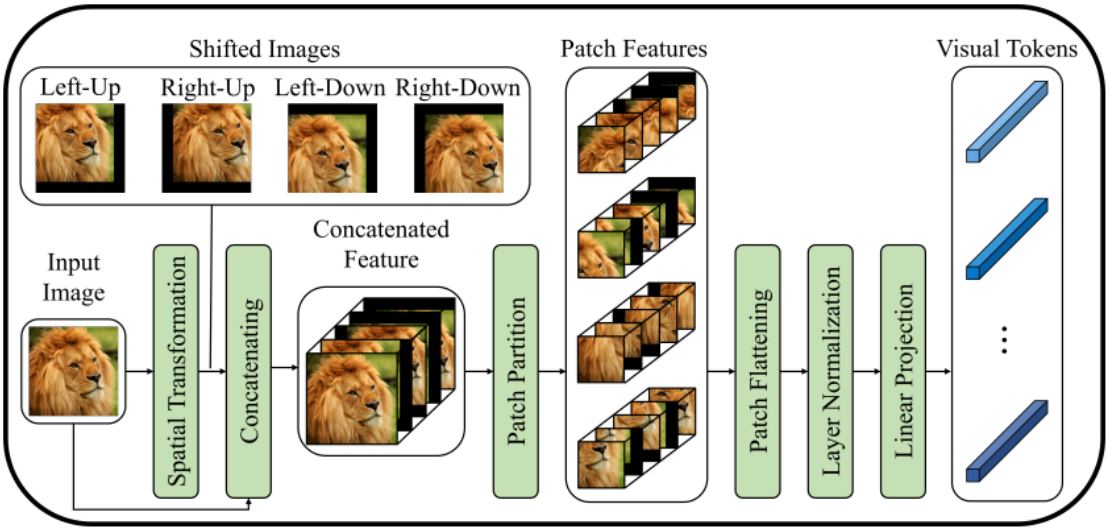

Figure 3: Shifted Patch Tokenization[3]

Description. The figure illustrates the Shifted Patch Tokenization (SPT) process. The input image is first shifted in four directions—Left-Up, Right-Up, Left-Down, and Right-Down—creating multiple spatially transformed versions. These shifted images are then concatenated with the original image before being divided into patches. Each patch is flattened, normalized, and passed through a linear projection layer, converting it into visual tokens for the Vision Transformer model.

are the only components that are translationally equivariant, as the same transformation is applied to all patches, regardless of their location (Dosovitskiy et al., 2020, p. 4). Due to this lack of inductive bias, training ViTs from scratch requires large datasets, or alternatively, pretrained models are needed to ensure effective learning (S. H. Lee et al., 2021, p. 1). Training the model from scratch with stock chart data is challenging in this case.

S. H. Lee et al. (2021) identify two key factors in Vision Transformers that contribute to their lack of locality inductive bias: poor tokenization and a weak attention mechanism for locality. In ViTs, images are divided into non-overlapping patches, meaning each patch is processed independently without considering neighboring pixels. In contrast, CNNs use sliding filters that capture not only the pixels in a specific region but also how surrounding pixels interact, increasing the receptive field and enhancing local feature extraction. Since ViTs treat each patch separately, they lack this spatial continuity, resulting in a weaker ability to encode fine-grained spatial relationships (S. H. Lee et al., 2021, p. 2). The second issue is the inefficiency of the attention mechanism when applied to image data. Unlike text-based Transformers, where the number of tokens is relatively small, ViTs must process hundreds of patches, significantly increasing the feature dimension and making self-attention computationally expensive. This leads to smooth attention scores, where attention is distributed too evenly across all patches instead of focusing on important regions. As a result, ViTs struggle to concentrate on critical parts of an image, often placing unnecessary attention on background details rather than the target object (S. H. Lee et al., 2021, p. 2).

To address these issues, (S. H. Lee et al., 2021) propose two solutions: Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA). SPT improves the locality of tokenization by shifting the input image in multiple directions before dividing it into patches. Figure 3 represents the shifting process of images and how they are transformed into visual tokens. This process embeds spatial relationships between neighboring pixels into the patch tokens, increasing local feature extraction. The shifted images are then concatenated with the original image, and standard patch partitioning is applied. By incorporating additional spatial information into the patch embeddings, SPT effectively expands the receptive field of ViTs, making them better at capturing local structures S. H. Lee et al. (2021, p. 3). Meanwhile, LSA enhances the attention mechanism by sharpening the attention score distribution. Instead of attention spreading too evenly across all patches, LSA learns a temperature parameter that controls the sharpness of the softmax function, allowing ViT to focus more on relevant regions. Additionally, diagonal masking is applied to remove self-token relations, ensuring that attention is directed towards meaningful patch interactions rather than being distributed uniformly S. H. Lee et al. (2021, p. 3). Together, SPT and LSA enhance the locality inductive bias of ViTs, improving their ability to recognize fine-g rained spatial details and focus on important image regions S. H. Lee et al. (2021, p. 1). I incorporated these two methods into the CaiT architecture, following (S. H. Lee et al., 2021), to enable CaiT to learn the stock image relationships from scratch.

4 Data and Methodology

In this paper, I replicate the study conducted by (Jiang et al., 2023) in their work "*(Re-)Imag(in)ing Price Trends*"

and extend it by replacing the Convolutional Neural Network with a Vision Transformer architecture, CaiT. Jiang et al. (2023) construct stock chart images from 5-day, 20-day, and 60-day periods to predict the subsequent 5-day, 20-day, and 60-day returns using CNN, resulting in a total of nine CNN models trained for different time horizons. Following their approach, this section describes the data used, the transformation process for generating stock chart images, and the modifications made to adapt these images for CaiT. I then provide an overview of the CNN and CaiT architectures, detailing their specifications and structures. Finally, I outline the training process, including optimization methods, loss functions, and techniques used to enhance model performance and generalization.

4.1 Data

Following (Jiang et al., 2023), the daily stock data is sourced from the Center for Research in Security Prices (CRSP) database and includes opening, high, low, and closing prices for stocks listed on the NYSE, AMEX, and NASDAQ exchanges along with volume and CRSP adjusted returns Jiang et al. (2023, p. 3205).

I adjusted the prices using the methodology of (Jiang et al., 2023) to represent price movements in a standardized format. First, data is structured into 5-day, 20-day, and 60-day periods, with each period consisting of consecutive trading days. The first closing price in each period is normalized to 1, and subsequent closing prices are iteratively calculated using CRSP-adjusted returns. Once the adjusted closing prices are determined, the opening, high, and low prices for each day are scaled proportionally based on their original relationship to that day's actual closing price Jiang et al. (2023, p. 3206). This step ensures price consistency across different stocks while also accounting for stock splits and dividends Jiang et al. (2023, p. 3199). Following (Jiang et al., 2023), moving average prices are calculated for each day in the period based on its length. For example, if 5-day images are created, moving average prices for the past 5 days are calculated Jiang et al. (2023, p. 3200).

Additionally, as described by (Jiang et al., 2023), if the return of any stock on the first day of a constructed 5-, 20-, or 60-day period is missing, it is excluded. Moreover, if the IPO or delisting of a stock occurs within a period, it is also excluded. However, if any other price or volume information is missing within those periods, the stock and period are included, but the missing values are left blank in the images Jiang et al. (2023, p. 3199).

4.2 Creating images of the stock charts

After processing the raw price data, I used the transformed data to generate the images following the methodology of (Jiang et al., 2023). The final image dataset consists of three different image sizes: 32×15, 64×60, and 96×180, corresponding to the 5-day, 20-day, and 60-day periods, respectively Jiang et al. (2023, p. 3202). The width of each image is always three times the period length, as each trading day spans three pixels. Within this structure, the first pixel column of each day represents the opening price, the second column represents the high and low prices, and the last column represents the closing price Jiang et al. (2023, p. 3199).

For the height of the price information, both the opening and closing prices are represented as single-pixel heights, while the high and low prices are marked with a single pixel at their respective levels, with the pixels in between filled to form a vertical bar Jiang et al. (2023, p. 3198). In addition to price data, the moving average line, derived from the calculated moving averages, is also included in the image. The price data and moving average line occupy the upper four-fifths of the image, while the lower part is reserved for volume bars, visually separating price trends from trading activity Jiang et al. (2023, p. 3198). The entire image is constructed on a black background, with all information—prices, moving averages, and volume bars—displayed in white Jiang et al. (2023, p. 3193). Figure 4 presents an example of the resulting 60-day images.

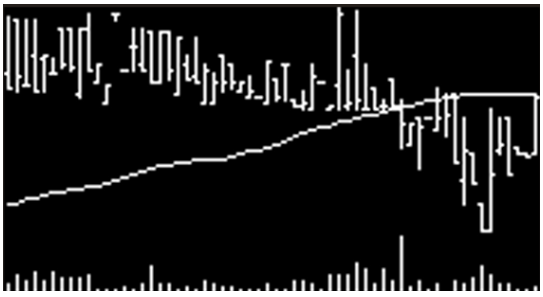

Figure 4: Example of 60-day Image for CNN

Description. This figure shows a 60-day stock chart image for CNN analysis. The lower part represents trading volume, while the upper part displays opening, high, low, and closing prices. The moving average line runs through the price section.

As described by (Jiang et al., 2023), in each stock chart, a consistent image height is ensured by scaling prices and volumes relative to their maximum and minimum values. For price and moving average lines, the values are rescaled so that their heights within the price section of the image fall between the maximum and minimum of either the opening, high, low, or close prices. Each price or moving average value is then adjusted according to this new height Jiang et al. (2023, p. 3198). Similarly, the volume bars are scaled within their designated area based on the maximum and minimum volume values Jiang et al. (2023, p. 3200). This scaling process standardizes stock price representations, making them comparable across different stocks and ensuring that they can be effectively processed by a Convolutional Neural Network Jiang et al. (2023, p. 3194).

I generated the images for CNN following the methodology of (Jiang et al., 2023); however, these images differ from those commonly used in Vision Transformer models, such as those in (Dosovitskiy et al., 2020; Touvron et al., 2021), and (S. H. Lee et al., 2021). The first key difference is that these Vision Transformer models use RGB images, meaning they

have three channels, whereas the images I created were originally grayscale, represented by a single channel. The second difference is in the image shape—standard ViT implementations use square patches, meaning each patch has equal width and height, while the images I created have rectangular shapes, which cannot be divided into square patches without distortion.

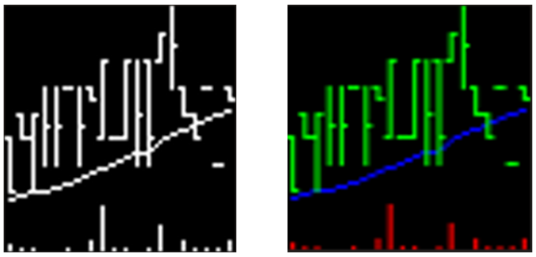

Figure 5: Grayscale and RGB Images of Stock Data

Description. The figure compares grayscale and RGB images of 20-day stock data. The left image (64×60) is a grayscale version used for CNN models, where stock prices, moving averages, and volume are represented in white on a black background. The right image (64×64) is the RGB-transformed and resized version for CaiT, with price data in green, the moving average line in blue, and volume bars in red, encoding different financial components into separate color channels.

To adapt the images for CaiT, I first converted them into RGB format, where each channel represents a specific type of information from the grayscale images. In this transformation, price data is represented in green, the moving average line in blue, and volume bars in red. Figure 5 shows an example of a 20-day stock image used for CNN (left) along with the RGB image (right) of the same stock. The original grayscale images I generated for CNNs are essentially matrices of the same shape as the image, where black pixels correspond to 0 and white pixels correspond to 255 Jiang et al. (2023, p. 8). In contrast, the RGB images used for CaiT consist of three separate matrices, each corresponding to a different color channel. These matrices are layered on top of each other, with red at the front, green in the middle, and blue at the back. In each channel matrix, the black areas are encoded as 0, while the colored regions are set to 255, effectively encoding different financial data components into separate channels for use in CaiT models.[4] Subsequently, following (Jiang et al., 2023), the pixel values of both grayscale and RGB images are normalized.

Secondly, to ensure compatibility with square patch partitioning in Vision Transformers, I transformed the images into a square format. Specifically, 32×15 images were resized to 32×32, and 64×60 images were adjusted to 64×64. This resizing was performed using bilinear interpolation, a method that smooths pixel values by averaging nearby pixels (Darji, 2021). The color variation in the right image of Figure 5 is a result of the resizing process. The transformation slightly alters pixel intensities, causing minor shifts in color representation. Although the original images maintained a fixed three-pixel width per trading day structure, this arrangement no longer applies explicitly after resizing. However, this information is still indirectly preserved, as bilinear interpolation was applied consistently across all images. Since every image in a given period was resized the same way, the resulting representations retain the encoded structure of the original format, ensuring that the attention mechanism in Vision Transformers can still capture meaningful relationships between patches. I opted for resizing instead of incorporating images with different time horizons into a single model to ensure a direct comparison between CNN and CaiT predictions. By applying the same preprocessing steps to both models, they learn from identical input data, ensuring that any observed performance differences arise from architectural distinctions rather than disparities in data structure.

4.3 CNN Architecture Specifications

The CNN architecture used in this paper is proposed by (Jiang et al., 2023) and designed to process stock price charts of different time horizons—5-day, 20-day, and 60-day—by progressively increasing network depth as the input size grows. For the 5-day input image with a resolution of 32×15, the network consists of two convolutional layers with 5×3 filters, where the first layer has 64 channels and the second has 128 channels. Each convolutional layer is followed by a 2×1 max-pooling layer, and the extracted features are passed through a fully connected layer of size 15,360 before classification via a softmax layer Jiang et al. (2023, p. 3202).

For the 20-day input image with a resolution of 64×60, the model becomes deeper with three convolutional layers, all using 5×3 filters, with the number of channels increasing from 64 to 128 to 256. Each convolutional layer is followed by a 2×1 max-pooling layer, reducing the spatial dimensions while preserving key features. The flattened feature representation is then passed through a fully connected layer of size 46,080, and classification is performed using a softmax function Jiang et al. (2023, p. 3202).

The 60-day input image, with a resolution of 96×180, is processed by an even deeper CNN. This version of the model consists of four convolutional layers with 5×3 filters, where the number of channels increases from 64 to 128, then to 256, and finally to 512. As in the previous models, each convolutional layer is followed by a 2×1 max-pooling layer, which progressively reduces dimensionality while preserving essential trend-related features. The extracted features are then passed through a fully connected layer of size 184,320, before classification with a softmax layer Jiang et al. (2023, p. 3202).

Following (Jiang et al., 2023), all versions of the CNN employ leaky ReLU, a variant of ReLU proposed by (Maas et al., 2013), as the activation function after each convolutional layer. This choice introduces non-linearity while mitigating the issue of dead neurons Jiang et al. (2023, p. 3239).

[4] In Figure 5, the price bars appear to intersect with and obscure parts of the moving average line. However, the moving average information is still encoded within its respective channel matrix; it is simply not visible in this representation.

The deeper architectures used for longer time horizons allow the model to capture increasingly complex patterns in stock price movements, with the max-pooling layers ensuring that important spatial relationships in the input images are preserved while reducing computational complexity (Jiang et al., 2023, pp. 3240–3242). The fully connected layers aggregate extracted features before the final classification stage, enabling the model to make predictions about future price movements based on historical chart patterns Jiang et al. (2023, p. 3241).

4.4 CaiT Architecture Specifications

Following (S. H. Lee et al., 2021), I used the CaiT architecture as proposed by (Touvron et al., 2021). The structure of CaiT is implemented as outlined by (Touvron et al., 2021) with two modifications: Locality Self-Attention and Shifted Patch Tokenization, as suggested by (S. H. Lee et al., 2021).

Following (Touvron et al., 2021), my CaiT model consists of 24 Transformer layers for processing patch embeddings, followed by 2 layers dedicated to class attention. Each multi-head self-attention layer has 4 attention heads, with each head having a dimension of 64, resulting in a total attention dimension of 256 per layer. The model uses a patch size of 4×4, ensuring fine-grained tokenization of the image. Each patch is projected into an embedding dimension of 192, serving as the feature representation for further processing. The MLP expansion ratio is set to 2, meaning the hidden dimension of the feedforward layers is 384. Similar to the CNN model, the CaiT model is designed for a binary classification task with two output classes, corresponding to predicting whether stock returns will go rise or fall.

Following (Touvron et al., 2021), within each Transformer layer, Layer Normalization (Ba et al., 2016) is applied before both the self-attention and MLP layers to maintain stable activations and ensure efficient learning. Residual connections (He et al., 2016b) are included in each Transformer layer, facilitating for better gradient flow and preventing vanishing gradients. To further enhance training stability in this deep architecture, LayerScale as proposed by (Touvron et al., 2021) is applied to residual connections. In MLP layers, GeLU (Hendrycks & Gimpel, 2016) activation function is used.

Following (S. H. Lee et al., 2021), Locality Self-Attention is applied to each self-attention mechanism in both Patch and Class Transformer layers, sharpening attention score distributions through a learnable temperature parameter. In the tokenization step, Shifted Patch Tokenization is used, where the input image is shifted in multiple directions before being divided into patches, ensuring that more spatial relationships are embedded into each visual token.

With these design choices, the model is optimized for learning complex spatial relationships within stock charts, allowing it to capture subtle patterns and improve classification performance. The model specifications or structure don't change with respect to the image size, as in CNN models.

4.5 Training

Now, I will describe the training process for both CNN and CaiT. For the CNN models, I follow the methodology outlined by (Jiang et al., 2023), which is itself based on the training procedure introduced by (Gu et al., 2020). The dataset spans from 1993 to 2000 for training, while the period from 2001 to 2019 is designated as the out-of-sample dataset for evaluation. Within the training dataset (1993–2000), the data is further split into 70% training and 30% validation sets, ensuring proper model tuning and preventing overfitting Jiang et al. (2023, p. 3204). These training and validation samples are selected randomly while maintaining a balanced classification task, accounting for both bullish and bearish market trends to avoid bias in model learning Jiang et al. (2023, p. 3204). I applied this same structured workflow to the CaiT models, allowing for a direct comparison between the CNN and Vision Transformer architectures in predicting stock price movements.

The training process begins by feeding the training dataset into the neural network, where the model makes predictions based on the input images. Once the model has been trained on this data, the validation set is passed through the network to evaluate its performance on unseen data Goodfellow et al. (2016, p. 204). Following (Jiang et al., 2023), the prediction task is formulated as a binary classification problem, where stock returns are categorized into two classes: up or down. Labels are assigned based on the actual return values, with positive returns labeled as 1 and negative returns labeled as 0 Jiang et al. (2023, p. 3204). The loss function, which the model attempts to minimize, is set to cross-entropy loss Jiang et al. (2023, p. 3205) in both CNN and CaiT.

Again, I followed the regularization techniques outlined by (Jiang et al., 2023), which are based on (Gu et al., 2020). To train the model, the Adam optimization algorithm (Kingma & Ba, 2017) is used, dynamically adjusting learning rates for efficient weight updates. The training begins with an initial learning rate of 1×10^{-5} and the batch size is 128 Jiang et al. (2023, p. 3205). To stabilize learning, batch normalization (Ioffe & Szegedy, 2015) is applied between convolutional layers and activation functions, ensuring a consistent distribution of activations and reducing internal covariate shift Jiang et al. (2023, p. 3205). In addition, dropout regularization (Srivastava et al., 2014) is used to further prevent overfitting, with 50% of neurons in the fully connected layer randomly deactivated during training. However, dropout is not applied to convolutional layers, as their lower parameter count naturally limits overfitting Jiang et al. (2023, p. 3205). For weight initialization, I use the Xavier initializer (Glorot & Bengio, 2010), which assigns initial weight values to facilitate faster convergence by maintaining balanced variance across layers. Furthermore, early stopping is implemented with a patience of two epochs, meaning training stops if the validation loss does not improve for two consecutive epochs, preventing excessive training and reducing the risk of overfitting Jiang et al. (2023, p. 3205). Finally, following (Gu et al., 2020) and

(Jiang et al., 2023), five independent models are trained, and their predictions are averaged to form the final output, ensuring a more stable and reliable classification result.

For the CaiT models, following Lee et al. (2022), I applied label smoothing (Szegedy et al., 2016) to the cross-entropy loss function. Label smoothing is a regularization technique that prevents the model from becoming overconfident in its predictions by slightly adjusting the true class probabilities. Instead of assigning a hard label of 0 or 1, it redistributes a small portion of the probability mass to all classes Goodfellow et al. (2016, p. 243). This technique helps reduce overfitting and improves generalization, particularly in deep networks where extreme confidence in predictions can lead to poor robustness Szegedy et al. (2016, p. 2823). Following (S. H. Lee et al., 2021), AdamW (Loshchilov & Hutter, 2019) optimizer is chosen with a weight decay of 0.05 and a warm-up period is set to 5. Stochastic depth (Huang et al., 2016) is also applied. For learning rate settings, I set the initial learning rate to 1×10^{-5} for 20-day images and to 1×10^{-4} for 5-day images, following a structured adjustment for different time horizons. Batch size is set to 128, similar to CNN. Although (S. H. Lee et al., 2021) propose several data augmentation techniques, I did not incorporate these methods in my study. Since my task involves stock charts rather than natural images, I was uncertain about their effectiveness for time-series-based graphs. Unlike object classification tasks where mixing pixels or cutting portions of an image may still preserve semantic meaning, altering stock chart images in such a manner could disrupt the underlying financial patterns crucial for prediction.

5 Results

In this section, I present the performance results of the models over the 19-year out-of-sample period. Following (Jiang et al., 2023), stocks are first sorted into deciles based on the probability outputs of their respective models. Then, the annualized average returns for each holding period, along with their annualized Sharpe ratios to assess risk-adjusted performance, are calculated. Additionally, long-short portfolios (H-L) are constructed by longing the highest decile (Decile 10 or High) and shorting the lowest decile (Decile 1 or Low) to examine the return spread between extreme predictions.

For each model name, I use a notation format where the number of days the images cover is followed by "D", and the prediction horizon for returns is followed by "R". For example, "20D/5R" indicates that 20-day images are used to predict the next 5-day returns. Following (Jiang et al., 2023), portfolios are rebalanced based on the prediction period, meaning models that predict 5-day, 20-day, and 60-day returns result in weekly, monthly, and quarterly rebalancing, respectively. Following (Jiang et al., 2023), I compute the monthly turnover for each model, as described by (Gu et al., 2020). For strategies rebalanced weekly or quarterly, turnover is appropriately scaled to a monthly basis. A 200% turnover indicates a complete replacement of stocks within one period, while 0% turnover means no changes were made Jiang et al. (2023, p. 17).

I begin by presenting the results for the CNN models, alongside comparisons with Momentum, Short-Term Reversal, and Weekly Short-Term Reversal strategies to provide additional benchmarks following (Jiang et al., 2023). Next, I report the results for the CaiT models, evaluating their predictive effectiveness in the same framework. Finally, I take the simple average of the probability outputs from both CNN and CaiT models, sort stocks based on these average probabilities and analyze the results. This combined approach, referred to as the "Average" model, offers insight into whether the integration of both architectures improves predictive accuracy and portfolio performance.

5.1 CNN Results

In this section, I discuss the results of CNN models that predict returns over 5-, 20-, and 60-day horizons, starting with the weekly rebalanced strategies. Following (Jiang et al., 2023), Table 1 focuses on short-horizon strategies 5D/5R and 20D/5R compared to traditional momentum MOM, monthly short-term reversal STR, and weekly short-term reversal WSTR. In the equal-weighted panel (top), the image-based models begin with negative Sharpe ratios in the lowest decile -1.91 for 5D/5R and -2.06 for 20D/5R but rise monotonically to 2.91 and 2.70 in the highest decile, underscoring a robust positive relationship between the CNN up probabilities and subsequent returns. Moreover, the long–short H–L portfolios deliver exceptionally high annualized Sharpe ratios of 7.20 for 5D/5R and 6.96 for 20D/5R, far exceeding the Sharpe ratios of 0.12, 1.75 and 2.91 for MOM, STR and WSTR, respectively. A similar pattern appears in the value-weighted panel (below), though at a lower level, where the H–L Sharpe ratios are 1.54 for 5D/5R and 1.64 for 20D/5R, again outperforming the benchmark signals by a wide margin. Finally, Table 1 indicates that these short-horizon image-based strategies involve relatively high monthly turnover. In the equal-weight panel, 5D/5R and 20D/5R each exceed 660% turnover, matching WSTR at 660% and contrasting with STR at 341% and MOM at 124%. The value-weighted panel follows a similar pattern, with the image-based strategies exceeding 700%, short-term reversal near 433%, and momentum at 164%.

At a monthly rebalancing frequency, image-based predictions continue to demonstrate their superiority over traditional signals. Following (Jiang et al., 2023), Table 2 depicts the monthly rebalanced portfolios using 5D/20R, 20D/20R and 60D/20R, again benchmarked against MOM, STR and WSTR. For equal-weighted portfolios, 5D/20R and 20D/20R yield H-L Sharpe ratios of 2.44 and 2.19, respectively, which is significantly higher than the highest traditional trend signal, WSTR, with the H-L Sharpe ratio of 1.29. The 60D/20R model, which relies on 60-day images, achieves a Sharpe ratio of 1.04. Traditional short-term reversal (STR) lags significantly with a Sharpe ratio of 0.54, and the momentum strategy performs the weakest at 0.32. Furthermore, the image-based strategies trade at relatively high turnover, roughly

Table 1: One Week Portfolio Performance

Equal Weight										
	5D/5R		20D/5R		MOM/5R		STR/5R		WSTR/5R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.27	-1.91	-0.33	-2.06	0.13	0.40	-0.01	-0.05	-0.09	-0.41
2	-0.05	-0.27	-0.04	-0.20	0.10	0.43	0.06	0.37	0.04	0.22
3	0.02	0.10	0.04	0.20	0.11	0.53	0.09	0.58	0.08	0.49
4	0.07	0.38	0.09	0.45	0.11	0.59	0.10	0.67	0.09	0.59
5	0.10	0.49	0.14	0.71	0.12	0.71	0.11	0.72	0.09	0.60
6	0.13	0.67	0.16	0.83	0.12	0.79	0.11	0.71	0.11	0.68
7	0.17	0.85	0.19	1.00	0.13	0.87	0.12	0.68	0.13	0.76
8	0.22	1.09	0.22	1.16	0.15	0.95	0.12	0.62	0.15	0.78
9	0.30	1.48	0.27	1.42	0.15	0.90	0.16	0.68	0.19	0.83
High	0.54	2.91	0.50	2.70	0.16	0.77	0.37	1.17	0.45	1.55
H-L	0.82	7.20	0.83	6.96	0.03	0.12	0.38	1.75	0.54	2.91
Turnover	681%		663%		124%		341%		660%	
Value Weight										
	5D/5R		20D/5R		MOM/5R		STR/5R		WSTR/5R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.03	-0.17	-0.05	-0.26	0.02	0.05	0.02	0.10	-0.05	-0.21
2	0.01	0.03	0.02	0.11	0.03	0.11	0.03	0.18	0.01	0.07
3	0.03	0.19	0.03	0.18	0.08	0.32	0.06	0.37	0.06	0.34
4	0.07	0.45	0.07	0.46	0.07	0.38	0.08	0.53	0.07	0.48
5	0.06	0.37	0.06	0.38	0.08	0.52	0.08	0.60	0.08	0.60
6	0.08	0.54	0.07	0.45	0.08	0.50	0.09	0.61	0.09	0.64
7	0.08	0.51	0.09	0.62	0.07	0.43	0.11	0.68	0.11	0.77
8	0.08	0.51	0.08	0.51	0.08	0.53	0.10	0.58	0.10	0.59
9	0.10	0.64	0.08	0.54	0.08	0.54	0.08	0.39	0.10	0.54
High	0.11	0.71	0.09	0.55	0.09	0.68	0.05	0.18	0.08	0.29
H-L	0.06	0.47	0.02	0.16	0.04	0.28	0.02	0.09	0.07	0.36
Turnover	188%		182%		164%		184%		185%	

Description. The table reports annualized average weekly returns (ret), annualized Sharpe ratios (SR), and monthly turnover for Equal-Weighted (top) and Value-Weighted (bottom) portfolios, comparing different models across deciles, with long-short (H-L) portfolios capturing return spreads. Modified, taken from Jiang et al. (2023, p. 3207).

ranging between 150% and 175%, close to STR turnover of 168% and WSTR turnover of 167%, while MOM trades at a turnover of 64%. Similar to weekly strategies, as we move up the decile rankings, the risk-adjusted returns increase consistently, showcasing a strong monotonic relationship between the CNN's predicted probability and actual returns. In the value-weighted context, the 5D/20R model stands out as the strongest performer with a H-L Sharpe ratio of 0.47, followed by MOM and WSTR with H-L Sharpe ratios of 0.36.

Following (Jiang et al., 2023), Table 3 shows the performance of quarterly rebalancing strategies, still outperforming traditional signals but only with one strategy. In the equal-weighted portfolios, 5D/60R gives the best H-L Sharpe ratio with 1.16, followed by WSTR with a Sharpe ratio of 0.69. 20D/60R and 60D/60R have Sharpe ratios of 0.45 and 0.40, respectively, falling behind the WSTR. MOM and STR give the least Sharpe ratios with 0.12 and 0.32, respectively. Monthly turnover for image-based strategies along with STR and WSTR falls between 57% and 60%, whereas MOM has a slightly lower turnover of 38%. 20D/60R and 60D/60R still show a monotonic increase going from the Low decile to High decile.

5.2 CaiT Results

In this section, the performance of CaiT predictions is analyzed in detail. Due to computational cost, only three CaiT models are trained to assess their effectiveness in stock return prediction. The first model utilizes 5-day images to predict 60-day returns. This model was selected because it has a more limited dataset compared to models predicting 5-day or 20-day returns, as it only considers stock data at 60-day intervals. Additionally, it employs the shortest image length (5-day images), encoding minimal information. This serves as an essential benchmark to evaluate the feasibility of applying Vision Transformers to stock classification tasks. If the performance of CaiT is comparable to that of CNN, despite the restricted data and limited encoded knowledge, it may suggest that CaiT could be effectively utilized for other image sizes and prediction windows.

The other two models predict 5-day returns, with one using 5-day images and the other using 20-day images. Vision Transformers employ a global attention mechanism to identify relationships between different regions of an image. Intuitively, using a longer image period (20 days) while analyzing the same stocks should enable the model to capture more meaningful relationships compared to shorter-period images (5 days). The objective of these two models is to assess how performance is influenced by incorporating additional historical data into the images of the same stocks. For comparison, CNN models with the same configurations were chosen as benchmarks rather than traditional price signals to highlight the differences between the two image recognition architectures. The same notation used for CNN models is maintained, with a prefix of "CaiT-" or "CNN-" to distinguish between the two. For example, "CaiT-5D/5R" represents a CaiT model that uses 5-day images to predict 5-day returns, whereas "CNN-5D/5R" denotes the corresponding CNN model.

Table 4 presents the portfolio performance of the CaiT models benchmarked against their corresponding CNN models. For the 5D/60R configuration, both models exhibit comparable performance. In equal-weighted portfolios, CaiT-5D/60R achieves a H-L portfolio return of 0.08 with a Sharpe ratio of 1.23, while CNN-5D/60R yields the same return with a slightly lower Sharpe ratio of 1.16. Additionally, the performance of the Low and High deciles is similar between the two models. However, CaiT-5D/60R demonstrates a more distinct monotonic increase from the Low to High decile, suggesting a stronger ability to differentiate between stock performance categories. Both models exhibit identical turnover rates. The similar performance could be attributed to the limited information contained in the small images and the Shifted Patch Tokenization mechanism described in Section 3.5, which may allow CaiT patches to function similarly to convolutional filters, resulting in the extraction of comparable local features.

For the 5D/5R configuration, the performance similarity between CaiT and CNN persists, with the main difference observed in the High decile. In equal-weighted portfolios, both models yield a Sharpe ratio of -1.91 in the Low decile with nearly identical returns. In the High decile, both models achieve a return of 0.54; however, the Sharpe ratio is slightly higher for CNN-5D/5R (2.91) compared to CaiT-5D/5R (2.78). Consequently, the H-L portfolio return remains consistent at 0.82, with Sharpe ratios of 7.06 for CaiT-5D/5R and 7.20 for CNN-5D/5R. This suggests that, for the same level of risk, CNN delivers slightly higher returns. Regarding turnover, CaiT-5D/5R exhibits a lower turnover rate (677%) compared to CNN-5D/5R (681%).

For the 20D/5R configuration, the performance dynamics shift. CaiT-20D/5R achieves an H-L Sharpe ratio of 7.20, matching that of CNN-5D/5R, while CNN-20D/5R records a slightly lower Sharpe ratio of 6.96. In the low decile, CNN-20D/5R has a Sharpe ratio of -2.06, indicating that it captures more negative returns per unit of risk compared to CaiT. In the high decile, both models yield a Sharpe ratio of approximately 2.7. However, this does not translate to a higher Sharpe ratio in the H-L portfolio. Despite having a lower H-L return of 0.75, CaiT-20D/5R is able to predict higher returns for the same level of risk as CNN. The turnover for CaiT-20D/5R remains slightly lower than that of CNN-20D/5R.

In conclusion, for equal-weighted portfolios, CaiT and CNN 5D/60R models produce similar returns and Sharpe ratios. In the 5D/5R configuration, both models exhibit comparable returns; however, CNN outperforms CaiT in terms of risk-adjusted return in the High decile, leading to a superior risk-adjusted return for the H-L portfolio. When larger images are utilized, the performance dynamic shifts in favor of CaiT. CaiT-20D/5R does not generate negative risk-adjusted returns to the same extent as CNN in the Low decile, and both models achieve equivalent risk-adjusted returns in the high decile. However, in the H-L portfolio, CaiT-20D/5R demonstrates a higher return for the same level of risk, highlighting

Table 2: One Month Portfolio Performance

Equal Weight												
	5D/20R		20D/20R		60D/20R		MOM/20R		STR/20R		WSTR/20R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.01	-0.05	-0.02	-0.09	0.00	-0.01	0.05	0.15	0.05	0.21	-0.01	-0.02
2	0.05	0.28	0.05	0.27	0.06	0.29	0.07	0.28	0.08	0.46	0.06	0.32
3	0.07	0.39	0.08	0.43	0.09	0.45	0.09	0.45	0.10	0.62	0.09	0.57
4	0.08	0.42	0.10	0.54	0.10	0.52	0.09	0.52	0.10	0.70	0.10	0.67
5	0.09	0.51	0.10	0.57	0.11	0.61	0.10	0.63	0.11	0.76	0.11	0.75
6	0.11	0.59	0.11	0.63	0.12	0.67	0.11	0.81	0.11	0.69	0.12	0.77
7	0.12	0.62	0.13	0.73	0.13	0.75	0.13	0.93	0.12	0.74	0.11	0.71
8	0.13	0.70	0.13	0.75	0.13	0.79	0.13	0.97	0.11	0.58	0.12	0.68
9	0.17	0.90	0.15	0.90	0.14	0.88	0.14	0.90	0.09	0.42	0.13	0.62
High	0.20	1.10	0.17	1.01	0.14	1.01	0.14	0.73	0.15	0.49	0.19	0.65
H-L	0.21	2.44	0.19	2.19	0.14	1.04	0.09	0.32	0.10	0.54	0.19	1.29
Turnover	176%		174%		152%		64%		168%		167%	
Value Weight												
	5D/20R		20D/20R		60D/20R		MOM/20R		STR/20R		WSTR/20R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	0.05	0.33	0.07	0.38	0.06	0.27	0.01	0.02	0.03	0.14	0.01	0.05
2	0.04	0.24	0.06	0.37	0.08	0.43	0.03	0.12	0.06	0.35	0.03	0.18
3	0.05	0.33	0.06	0.35	0.05	0.28	0.05	0.24	0.06	0.38	0.05	0.35
4	0.07	0.45	0.07	0.46	0.07	0.38	0.07	0.38	0.08	0.53	0.07	0.48
5	0.07	0.36	0.07	0.39	0.10	0.54	0.08	0.45	0.10	0.63	0.10	0.62
6	0.09	0.48	0.08	0.39	0.07	0.39	0.10	0.64	0.09	0.53	0.08	0.49
7	0.08	0.44	0.07	0.37	0.06	0.33	0.08	0.57	0.10	0.58	0.09	0.57
8	0.10	0.56	0.07	0.43	0.07	0.39	0.07	0.50	0.09	0.46	0.09	0.46
9	0.08	0.53	0.09	0.56	0.08	0.51	0.08	0.51	0.08	0.35	0.09	0.43
High	0.10	0.58	0.10	0.64	0.09	0.55	0.10	0.48	0.07	0.21	0.06	0.22
H-L	0.04	0.44	0.04	0.36	0.02	0.14	0.03	0.09	0.03	0.13	0.05	0.24
Turnover	62%		60%		59%		43%		61%		62%	

Description. The table reports annualized average monthly returns (ret), annualized Sharpe ratios (SR), and monthly turnover for Equal-Weighted (top) and Value-Weighted (bottom) portfolios, comparing different models across deciles, with long-short (H-L) portfolios capturing return spreads. Modified, taken from Jiang et al. (2023, p. 3207).

Table 3: One Quarter Portfolio Performance

Equal Weight												
	5D/60R		20D/60R		60D/60R		MOM/60R		STR/60R		WSTR/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	0.06	0.25	0.06	0.22	0.06	0.21	0.07	0.17	0.05	0.19	0.04	0.13
2	0.08	0.34	0.08	0.32	0.08	0.30	0.09	0.28	0.09	0.39	0.09	0.38
3	0.09	0.38	0.09	0.36	0.10	0.39	0.11	0.43	0.11	0.53	0.10	0.46
4	0.09	0.41	0.11	0.45	0.10	0.45	0.11	0.51	0.10	0.54	0.10	0.53
5	0.11	0.47	0.11	0.47	0.11	0.46	0.10	0.55	0.11	0.62	0.11	0.63
6	0.10	0.46	0.11	0.48	0.10	0.45	0.11	0.63	0.11	0.58	0.11	0.60
7	0.10	0.47	0.11	0.51	0.11	0.53	0.11	0.61	0.12	0.61	0.11	0.54
8	0.13	0.57	0.12	0.55	0.12	0.56	0.11	0.65	0.11	0.47	0.11	0.50
9	0.11	0.52	0.12	0.59	0.12	0.62	0.12	0.63	0.10	0.36	0.11	0.43
High	0.14	0.66	0.12	0.64	0.12	0.69	0.11	0.47	0.12	0.33	0.14	0.40
H-L	0.08	1.16	0.06	0.45	0.06	0.40	0.04	0.12	0.07	0.32	0.10	0.69
Turnover	60%		59%		58%		38%		57%		57%	
Value Weight												
	5D/60R		20D/60R		60D/60R		MOM/60R		STR/60R		WSTR/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	0.05	0.27	0.05	0.24	0.07	0.29	0.07	0.15	0.04	0.13	0.02	0.06
2	0.06	0.30	0.06	0.29	0.07	0.34	0.03	0.10	0.07	0.36	0.05	0.25
3	0.06	0.34	0.07	0.33	0.09	0.41	0.09	0.35	0.07	0.39	0.07	0.39
4	0.07	0.38	0.10	0.49	0.08	0.40	0.08	0.38	0.07	0.45	0.09	0.55
5	0.07	0.36	0.07	0.39	0.10	0.54	0.08	0.45	0.10	0.63	0.10	0.62
6	0.09	0.48	0.08	0.39	0.07	0.39	0.10	0.64	0.09	0.53	0.08	0.49
7	0.08	0.44	0.07	0.37	0.06	0.33	0.08	0.57	0.10	0.58	0.09	0.57
8	0.10	0.56	0.07	0.43	0.07	0.39	0.07	0.50	0.09	0.46	0.09	0.46
9	0.08	0.53	0.09	0.56	0.08	0.51	0.08	0.51	0.08	0.35	0.09	0.43
High	0.10	0.58	0.10	0.64	0.09	0.55	0.10	0.48	0.07	0.21	0.06	0.22
H-L	0.04	0.44	0.04	0.36	0.02	0.14	0.03	0.09	0.03	0.13	0.05	0.24
Turnover	62%		60%		59%		43%		61%		62%	

Description. The table reports annualized average quarterly returns (ret), annualized Sharpe ratios (SR), and monthly turnover for Equal-Weighted (top) and Value-Weighted (bottom) portfolios, comparing different models across deciles, with long-short (H-L) portfolios capturing return spreads. Modified, taken from Jiang et al. (2023, p. 3207).

Table 4: Comparison of CaiT and CNN-based Portfolio Performance

Equal Weight												
	CaiT Only						CNN Only					
	5D/5R		20D/5R		5D/60R		5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.28	-1.91	-0.29	-1.69	0.05	0.22	-0.27	-1.91	-0.33	-2.06	0.06	0.25
2	-0.03	-0.18	-0.01	-0.03	0.08	0.34	-0.05	-0.27	-0.04	-0.20	0.08	0.34
3	0.03	0.18	0.06	0.32	0.09	0.38	0.02	0.10	0.04	0.20	0.09	0.38
4	0.07	0.38	0.11	0.55	0.10	0.44	0.07	0.38	0.09	0.45	0.09	0.41
5	0.11	0.56	0.13	0.67	0.11	0.46	0.10	0.49	0.14	0.71	0.11	0.47
6	0.13	0.64	0.15	0.80	0.11	0.49	0.13	0.67	0.16	0.83	0.10	0.46
7	0.17	0.87	0.18	0.92	0.11	0.50	0.17	0.85	0.19	1.00	0.10	0.47
8	0.21	1.06	0.20	1.08	0.11	0.51	0.22	1.09	0.22	1.16	0.13	0.57
9	0.29	1.43	0.25	1.35	0.12	0.55	0.30	1.48	0.27	1.42	0.11	0.52
High	0.54	2.78	0.46	2.69	0.14	0.63	0.54	2.91	0.50	2.70	0.14	0.66
H-L	0.82	7.06	0.75	7.20	0.08	1.23	0.82	7.20	0.83	6.96	0.08	1.16
Turnover	677%		636%		60%		681%		663%		60%	
Value Weight												
	CaiT Only						CNN Only					
	5D/5R		20D/5R		5D/60R		5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.02	-0.13	-0.02	-0.12	0.05	0.25	-0.03	-0.17	-0.05	-0.26	0.05	0.27
2	0.03	0.15	0.04	0.20	0.08	0.38	0.01	0.03	0.02	0.11	0.06	0.30
3	0.05	0.28	0.04	0.23	0.07	0.36	0.03	0.19	0.03	0.18	0.06	0.34
4	0.04	0.24	0.06	0.35	0.06	0.33	0.06	0.33	0.05	0.28	0.07	0.38
5	0.07	0.40	0.07	0.40	0.08	0.40	0.08	0.43	0.07	0.38	0.07	0.36
6	0.07	0.42	0.07	0.37	0.07	0.38	0.06	0.36	0.07	0.39	0.09	0.48
7	0.09	0.52	0.09	0.48	0.08	0.43	0.12	0.63	0.10	0.59	0.08	0.44
8	0.11	0.58	0.10	0.58	0.08	0.46	0.14	0.74	0.11	0.59	0.10	0.56
9	0.14	0.76	0.12	0.66	0.10	0.58	0.16	0.78	0.12	0.68	0.08	0.53
High	0.21	1.00	0.15	0.87	0.10	0.59	0.21	0.95	0.17	0.88	0.10	0.58
H-L	0.23	1.64	0.17	1.48	0.05	0.51	0.24	1.54	0.22	1.64	0.04	0.44
Turnover	760%		712%		61%		758%		731%		62%	

Description. The table compares CaiT-based and CNN-based portfolio strategies, with the left side showing CaiT results and the right side showing CNN results for both equal-weighted and value-weighted portfolios. It presents annualized average period returns (ret) and annualized Sharpe ratios (SR) across decile portfolios. The High-Low (H-L) spread represents the performance difference between the highest and lowest deciles. Monthly turnover rates are also reported at the bottom.

its potential advantages when incorporating more extensive historical data.

Table 5 presents the transaction cost-adjusted performance of equal-weighted CNN-based portfolios, reporting returns and Sharpe ratios for the Low and High deciles, as well as the H-L portfolio, across different prediction horizons and image sizes. The two highest Sharpe ratios, 4.28 for 5D/5R and 4.25 for 20D/5R, correspond to the strategies that also exhibited the highest Sharpe ratios before transaction costs, indicating that short-term strategies remain robust even after accounting for trading costs. Among monthly strategies, the highest Sharpe ratio of 1.61 is observed for 5D/20R, while for quarterly strategies, 5D/60R achieves the best Sharpe ratio of 0.81. These results confirm that CNN-based strategies withstand transaction costs and continue to generate positive returns and Sharpe ratios.

5.3 Transaction Costs

In Sections 5.1 and 5.2, I demonstrated that both CNN- and CaiT-based strategies generate high Sharpe ratios alongside high turnovers. In this section, I present the prediction results after incorporating transaction costs into portfolio construction. To account for these costs, I follow the approach of (Jiang et al., 2023), which is based on the methodology outlined by (Ke et al., 2021). Ke et al. (2021) assume that each asset incurs a daily transaction cost of 10 basis points (bps) for large stocks (those exceeding the 80th percentile of the NYSE size distribution) and 20 bps for smaller stocks (those below the 80th percentile; Ke et al. (2021, p. 31). This approach is motivated by (Frazzini et al., 2018), who analyze $1.7 trillion in live institutional trades and find average trading costs of 9.93 bps for large stocks and 20.18 bps for small stocks Frazzini et al. (2018, p. 53). Therefore, I adopt these transaction cost adjustments in my replication to ensure consistency with (Jiang et al., 2023) and prior literature on empirical trading cost estimation.

Table 5 presents the transaction cost-adjusted performance of equal-weighted CNN-based portfolios, reporting returns and Sharpe ratios for the Low and High deciles, as well as the H-L portfolio, across different prediction horizons and image sizes. The two highest Sharpe ratios, 4.28 for 5D/5R and 4.25 for 20D/5R, correspond to the strategies that also exhibited the highest Sharpe ratios before transaction costs, indicating that short-term strategies remain robust even after accounting for trading costs. Among monthly strategies, the highest Sharpe ratio of 1.61 is observed for 5D/20R, while for quarterly strategies, 5D/60R achieves the best Sharpe ratio of 0.81. These results confirm that CNN-based strategies withstand transaction costs and continue to generate positive returns and Sharpe ratios.

Table 6 presents the transaction cost-adjusted performance of CaiT-based strategies, structured in the same format as Table 5 for CNN. The results show that all CaiT-based strategies survive transaction costs, continuing to generate positive returns and Sharpe ratios. CaiT-5D/5R has a Sharpe ratio of 4.18, and CaiT-20D/5R has a Sharpe ratio of 4.05, both of which are slightly lower than their CNN counterparts. In contrast, CaiT-5D/60R strategy achieves a Sharpe ratio of 0.87, which is slightly higher than the CNN-based 5D/60R.

5.4 Average model results

As previously explained, CNN models generate probability estimates for individual stocks at each date, indicating the likelihood of a positive return. Similarly, CaiT models compute the same probability for the same stocks on the same dates, making a direct comparison between their outputs possible. Due to this similarity, their probabilities can also be combined. For CNN models, five separate models are trained on the same dataset, and their resulting probabilities are averaged for each stock at each date. These final averaged probabilities are then used to construct portfolios throughout the paper. Following the same approach, I take the simple average of the final CNN probabilities (average of 5 models) and the CaiT probabilities to create a new set of probabilities. In this section, I present the performance of portfolios constructed using these simple average probabilities. To ensure clarity, and consistency with Section 5.2, I prefix these models with "Average-" to distinguish them from standalone CNN and CaiT results. Additionally, I compare these "Average" results with CNN-based portfolios to assess the impact of integrating CaiT probabilities with CNN outputs.

The objective of this analysis is to assess the level of agreement between the CNN and CaiT models, particularly in the extreme decile portfolios. To illustrate this concept, consider a simple example with three stocks: A, B, and C. Suppose the CNN model estimates the probabilities of a positive return as 0.5 for A, 0.6 for B, and 0.9 for C, while the CaiT model assigns probabilities of 0.9 for A, 0.7 for B, and 0.8 for C. If we compute the average probabilities for each stock, we obtain 0.7 for A, 0.65 for B, and 0.85 for C. When ranked, Stock C remains the highest, as both models assign it a relatively high probability, suggesting strong agreement. Stock A ranks second, as it receives a high probability from at least one model. Stock B ranks lowest, since it does not receive consistently high probabilities from either model.

Extending this logic to the full stock universe, at each date, stocks that receive high probabilities from both models are expected to be positioned in the higher deciles, as decile portfolios are sorted based on the average probabilities of the stocks. Stocks that receive a high probability from one model but only a moderate probability from the other are likely to follow. Similarly, stocks assigned low probabilities by both models will be concentrated in the lower deciles. By averaging the probabilities from the two models, this approach provides a framework to evaluate how the degree of agreement between CNN and CaiT affects decile portfolio performance. Agreement in this context refers to both models assigning consistently high or low probabilities to the same stocks. Therefore, this analysis will focus primarily on

[5] modified, taken from Jiang et al. (2023, p. 3207)

Table 5: Performance of CNN-based Strategies After Transaction Costs[5]

	5D/5R		20D/5R			
	Ret	SR	Ret	SR		
Low	-0.10	-0.70	-0.16	-0.98		
High	0.38	2.04	0.35	1.88		
H-L	0.48	4.28	0.51	4.25		
Turnover	681%		663%			
	5D/20R		20D/20R		60D/20R	
	Ret	SR	Ret	SR	Ret	SR
Low	0.03	0.19	0.02	0.10	0.04	0.16
High	0.17	0.91	0.14	0.83	0.11	0.80
H-L	0.14	1.61	0.12	1.38	0.07	0.53
Turnover	176%		174%		152%	
	5D/60R		20D/60R		60D/60R	
	Ret	SR	Ret	SR	Ret	SR
Low	0.07	0.30	0.07	0.27	0.07	0.26
High	0.13	0.61	0.11	0.59	0.11	0.63
H-L	0.06	0.81	0.04	0.27	0.04	0.25
Turnover	60%		59%		58%	

Description. The table shows the performance of CNN-based strategies after transaction costs, with weekly strategies at the top, monthly in the middle, and quarterly at the bottom. It reports returns (Ret), Sharpe ratios (SR), for High, Low and H-L portfolios for each configuration. Monthly turnover rates are also reported.

Table 6: Performance of CaiT-based Strategies After Transaction Costs

	5D/5R		20D/5R		5D/60R	
	Ret	SR	Ret	SR	Ret	SR
Low	-0.10	-0.68	-0.11	-0.63	0.07	0.27
High	0.38	1.96	0.31	1.82	0.12	0.58
H-L	0.48	4.18	0.42	4.05	0.06	0.87
Turnover	677%		636%		60%	

Description. The table shows the performance of CaiT-based strategies after transaction costs. It reports returns (Ret), Sharpe ratios (SR), for High, Low and H-L portfolios for each configuration. Monthly turnover rates are also reported.

the High and Low deciles, as well as the H-L decile, to determine whether stocks with extreme probabilities in both models exhibit stronger predictive power compared to CNN, which relies on a single probability estimate.

Table 7 presents the decile portfolio performance of the Average model for 5D/5R, 20D/5R, and 5D/60R, comparing it with the respective CNN-only results. The findings indicate that, in both equal-weighted and value-weighted portfolios, the Average model consistently exhibits slightly higher H-L returns and Sharpe ratios compared to CNN-only portfolios.

For Average-5D/5R in equal-weighted portfolios, the H-L portfolio achieves a return of 0.86 and a Sharpe ratio of 7.25, slightly outperforming CNN-5D/5R, which records an H-L return of 0.82 and a Sharpe ratio of 7.20. The primary differences between the Average model and CNN-only portfolios are concentrated in the Low and High deciles, where the Average model tends to improve the performance of High decile stocks while slightly reducing the performance of Low decile stocks.

A similar trend is observed in Average-20D/5R, where the H-L return and Sharpe ratio are marginally higher than those of CNN-only portfolios. The differences are again most pronounced in the extreme deciles, with the Low decile portfolio showing a slightly lower return and Sharpe ratio, while the High decile portfolio exhibits a marginally stronger performance compared to CNN-only results.

For Average-5D/60R, the H-L return and Sharpe ratio are again slightly higher than those of CNN, although the differences are minimal. Within the High decile portfolio, the returns and Sharpe ratios remain identical to CNN-only results, whereas the Low decile portfolio exhibits a slightly lower Sharpe ratio.

To assess whether the observed differences between the CNN-only H-L portfolios and the Average H-L portfolios are statistically significant or merely due to randomness, I conducted a paired t-test for 5D/60R, 5D/5R, and 20D/5R. The results indicate varying levels of statistical significance across different prediction horizons.

For 5D/60R, the t-statistic of 1.19 and p-value of 0.24 suggest that the difference between the CNN-only and Average

Table 7: Comparison of Average Model and CNN-based Portfolio Performance

Equal Weight												
	Average model						CNN only					
	5D/5R		20D/5R		5D/60R		5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.30	-2.08	-0.35	-2.14	0.05	0.22	-0.27	-1.91	-0.33	-2.06	0.06	0.25
2	-0.04	-0.24	-0.03	-0.16	0.07	0.32	-0.05	-0.27	-0.04	-0.20	0.08	0.34
3	0.02	0.11	0.04	0.23	0.09	0.39	0.02	0.10	0.04	0.20	0.09	0.38
4	0.07	0.40	0.11	0.55	0.10	0.43	0.07	0.38	0.09	0.45	0.09	0.41
5	0.10	0.50	0.13	0.68	0.10	0.42	0.10	0.49	0.14	0.71	0.11	0.47
6	0.14	0.69	0.16	0.83	0.11	0.51	0.13	0.67	0.16	0.83	0.10	0.46
7	0.17	0.83	0.18	0.96	0.11	0.50	0.17	0.85	0.19	1.00	0.10	0.47
8	0.22	1.09	0.22	1.14	0.12	0.54	0.22	1.09	0.22	1.16	0.13	0.57
9	0.30	1.46	0.27	1.42	0.12	0.54	0.30	1.48	0.27	1.42	0.11	0.52
High	0.57	2.96	0.51	2.84	0.14	0.66	0.54	2.91	0.50	2.70	0.14	0.66
H-L	0.86	7.25	0.86	7.31	0.09	1.21	0.82	7.20	0.83	6.96	0.08	1.16
Turnover	674%		648%		60%		681%		663%		60%	
Value Weight												
	Average model						CNN only					
	5D/5R		20D/5R		5D/60R		5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR	ret	SR
Low	-0.03	-0.19	-0.06	-0.36	0.05	0.25	-0.03	-0.17	-0.05	-0.26	0.05	0.27
2	0.00	0.02	0.01	0.06	0.05	0.25	0.01	0.03	0.02	0.11	0.06	0.30
3	0.03	0.18	0.04	0.19	0.07	0.37	0.03	0.19	0.03	0.18	0.06	0.34
4	0.05	0.29	0.05	0.30	0.08	0.41	0.06	0.33	0.05	0.28	0.07	0.38
5	0.07	0.41	0.06	0.32	0.07	0.38	0.08	0.43	0.07	0.38	0.07	0.36
6	0.07	0.40	0.06	0.35	0.08	0.40	0.06	0.36	0.07	0.39	0.09	0.48
7	0.10	0.54	0.09	0.51	0.09	0.47	0.12	0.63	0.10	0.59	0.08	0.44
8	0.12	0.66	0.11	0.63	0.08	0.50	0.14	0.74	0.11	0.59	0.10	0.56
9	0.17	0.85	0.13	0.70	0.09	0.52	0.16	0.78	0.12	0.68	0.08	0.53
High	0.22	1.01	0.17	0.90	0.10	0.61	0.21	0.95	0.17	0.88	0.10	0.58
H-L	0.25	1.63	0.23	1.79	0.05	0.52	0.24	1.54	0.22	1.64	0.04	0.44
Turnover	759%		726%		62%		758%		731%		62%	

Description. This table compares the performance of the Average model and CNN-based portfolios for equal-weighted and value-weighted strategies across different configurations. It reports average period returns (ret), Sharpe ratios (SR), High-Low (H-L) portfolio, and turnover rates.

H-L portfolios is not statistically significant at conventional levels. This implies that incorporating CaiT probabilities does not lead to a meaningful improvement over CNN for 5D/60R.

In contrast, for 5D/5R, the results are highly significant, with a t-statistic of 8.33 and a p-value of 2.72×10^{-16}, indicating a strong rejection of the null hypothesis. This suggests that averaging CNN and CaiT probabilities leads to a statistically significant improvement in portfolio performance, reinforcing the effectiveness of the Average model in short-term prediction horizons. Similarly, for 20D/5R, the t-statistic of 4.43 and p-value of 1.05×10^{-5} confirm that the Average model significantly outperforms the CNN-only model, albeit with slightly lower statistical strength than in the 5D/5R case.

To conclude, I took the simple average of CNN and CaiT probability estimates and sorted them into deciles. This approach does not necessarily provide insight into the effect of CaiT probabilities in the middle deciles, as changes in decile portfolio performance cannot be directly attributed to either CaiT or CNN. However, for the High and Low deciles, the impact of CaiT is more observable because stocks have only two possible directions—remaining in the decile or being replaced. At each portfolio formation date, CNN assigns certain stocks to the High and Low deciles. In the High decile, when probabilities are averaged, CaiT either confirms the selection by also assigning high probabilities or replaces some stocks with new ones. The same logic applies to the Low decile, where CaiT may either reinforce the selection of low-probability stocks or introduce new ones. Based on this, we can hypothesize that for Average-5D/5R and Average-20D/5R, the higher returns and Sharpe ratios compared to CNN-only portfolios suggest that the stocks introduced by CaiT have higher average returns than the stocks it removes. Similarly, in the Low decile, the reverse occurs, where the stocks CaiT replaces tend to have lower returns than those originally selected by CNN.

This effect also applies to CaiT itself. When comparing the results from Section 5.2, we observe that the H-L portfolio of the Average Model outperforms the CaiT-only portfolio, indicating that adding CNN probabilities has a positive impact on the returns and Sharpe ratio of the CaiT-only H-L portfolio.

To summarize, averaging CNN and CaiT probabilities results in H-L returns that are slightly higher than those of CNN-only portfolios. While the differences in returns are small, for 5D/5R and 20D/5R, these differences are highly statistically significant.

6 Investment strategy

In this section, I introduce the investment strategy developed using both CNN and CaiT predictions. The strategy involves going long on stocks that both models identify as most likely to increase in value and shorting stocks that both models predict as most likely to decline. In Section 6.1, I will discuss the investment thesis, explaining the rationale behind this strategy and how it was formulated based on the findings from previous sections, particularly how the agreement between the two models strengthens stock selection and enhances portfolio construction. In Section 6.2, I will evaluate the out-of-sample performance of the strategy, analyzing its effectiveness beyond the training period and assessing its robustness in real-world market conditions.

6.1 Investment Thesis

In Section 5.2, I compared CaiT-only results with CNN-only results. For 5D/5R, both models had the same H-L returns, but CNN had a higher Sharpe ratio. For 20D/5R, CaiT had lower H-L returns but a higher Sharpe ratio. These indicate that CNN and CaiT capture and leverage different characteristics when analyzing the same stock data. In Section 5.4, I averaged the probability estimates of both models and sorted stocks into deciles based on these average probabilities. The goal was to assess how the portfolios performed when there was some degree of agreement between the models, though this agreement was only implicitly defined. By taking the simple average, agreement was inferred by stocks with high probabilities from both models being placed in higher deciles, while stocks with low probabilities from both models were sorted into lower deciles. The results showed that Average models consistently produced higher H-L returns and Sharpe ratios than CNN-only results across all three model configurations. This suggests that combining both models' probability estimates has the potential to improve portfolio performance. To summarize, my results show that CNN and CaiT construct portfolios differently, and combining their outputs can improve performance. To achieve this, I will integrate both models' outputs by selecting stocks where both models strongly agree. However, directly using the probability estimates from both models is not ideal for two key reasons.

The first reason is label smoothing (Szegedy et al., 2016), a regularization technique applied in the CaiT architecture. As explained in Section 4.5, both models optimize their probability estimates by minimizing the same loss function, cross-entropy. However, in the CaiT model, label smoothing redistributes a small portion of the predicted probability mass from the most confident class to the other class, preventing overconfidence and improving model convergence. This means that CaiT outputs are inherently less extreme than CNN outputs, as label smoothing systematically reduces the probability assigned to the predicted class. Because of this regularization, the probability estimates produced by CaiT are not directly comparable to those from CNN, which does not use label smoothing. While CNN may generate more extreme probabilities, CaiT produces smoother, more calibrated probabilities. As a result, directly averaging or combining these probabilities without considering this difference could introduce bias in portfolio construction. Therefore, an alternative approach is necessary to properly integrate both models' outputs while accounting for this systematic difference in probability distributions.

The second reason why directly using the probability estimates from both models is not ideal is related to the concept

of confidence in these probabilities. Both CaiT and CNN architectures output two probability estimates: one representing the likelihood of the stock going up and the other representing the likelihood of the stock going down. Since these two probabilities sum to one, the structure inherently makes 0.5 the weakest confidence signal. If a model assigns a 0.5 probability for a stock to go up, it also assigns 0.5 probability for it to go down, indicating no clear preference and a high degree of uncertainty in the prediction. In contrast, a higher up probability, such as 0.7, represents a stronger confidence that the stock will increase in value compared to 0.6, just as a lower up probability, such as 0.3, represents stronger confidence in a decline compared to 0.4. This is because a 0.3 up probability corresponds to a 0.7 down probability, whereas a 0.4 up probability corresponds to a 0.6 down probability.

Therefore, directly using probability estimates without considering confidence means incorporating weak signals into the decision-making process. Throughout this paper, as well as in (Jiang et al., 2023), portfolio construction is based on sorting stocks into deciles using up probabilities. Under this approach, a stock with an up probability of 0.5 would be ranked above a stock with an up probability of 0.1, despite the fact that 0.5 actually represents maximum uncertainty rather than confidence in an upward movement. This misinterpretation of weak signals can affect portfolio performance. To better understand this impact, I conducted an analysis examining how portfolio performance changes when only more confident signals are used.

Chen et al. (2016) applied CNNs for time series analysis of the Taiwan Stock Index Futures, setting a confidence threshold of 0.65, where only predictions exceeding this threshold were considered valid for classification, while those below were discarded due to insufficient confidence Chen et al. (2016, p. 89). Inspired by this approach, I also incorporate confidence-based filtering but, instead of applying a fixed threshold, I use percentile-based filtering to observe how portfolio performance changes gradually as less confident signals are removed.

To achieve this, I first analyze the entire probability distribution for 5D/60R, 5D/5R, and 20D/5R across the out-of-sample period so there is a look-ahead bias in these results. I then iteratively remove probabilities between specific percentiles, starting with the 45th to 55th percentile, then expanding the exclusion range to 40th to 60th percentile, continuing this process until only the most confident signals—those in the 5th and 95th percentiles—remain. For each filtered dataset, I sort stocks into deciles using both CNN probabilities and the average of CNN and CaiT probabilities. This allows me to first examine how restricting the dataset to more confident CNN signals affect portfolio performance, and then evaluate the impact of incorporating CaiT probabilities into these subsets of confident CNN signals.

Figure 6 visualizes the H-L returns and Sharpe ratios for 5D/60R, 5D/5R, and 20D/5R under different percentile filtering thresholds. The figure also compares CNN portfolios with those constructed using average CNN and CaiT probabilities, providing insights into how increasing confidence thresholds impact portfolio returns and whether integrating CaiT probabilities enhances performance across varying confidence levels.

For 5D/60R, CNN H-L returns start at 0.08 and exhibit a gradual increase up to 0.13 at the 15th-85th percentile filter. However, after this point, returns decline slightly, ending at 0.125 in the 5th-95th percentile. In contrast, when CNN probabilities are averaged with CaiT probabilities, returns increase more consistently at each filtering step, starting at 0.085 and reaching 0.16 at the 5th-95th percentile. Regarding Sharpe ratios, CNN results do not show a clear upward trend, with most portfolios falling below the initial Sharpe ratio, except for the jump at 15th-85th. Meanwhile, the Average model's Sharpe ratios are mostly above the starting value and consistently outperform their CNN-only counterparts, though there is a slight decline at the final filtering level.

For 5D/5R and 20D/5R, the trend is notably different. Both exhibit a clear and consistent increase in H-L returns and Sharpe ratios for CNN and the Average model. The Average model consistently outperforms CNN portfolios at every filtering level. For 5D/5R, CNN starts with an H-L return of 0.82 and a Sharpe ratio of 7.2, which gradually increases to 1.7 in returns and 9.2 in Sharpe ratio. The Average model starts at 0.86 with a Sharpe ratio of 7.25 and increases to 1.90 in returns and 9.92 in Sharpe ratio, maintaining higher values than CNN-only portfolios at each filtering step. A similar pattern is observed for 20D/5R, where CNN begins with an H-L return of 0.83 and the Average model starts at 0.86. As filtering progresses, their returns increase to 1.99 and 2.17, respectively, while their Sharpe ratios rise from 6.95 and 7.30 to 9.19 and 9.98, respectively.

These findings are important because they demonstrate that gradually filtering out weaker probability signals leads to improved H-L returns and Sharpe ratios. Additionally, the Average model consistently outperforms CNN, reinforcing the value of integrating CaiT predictions. This further supports the hypothesis that removing uncertain predictions enhances model effectiveness and that combining CNN and CaiT probabilities produces superior portfolio performance across various confidence thresholds.

Based on these findings, I propose an investment strategy that selectively integrates confident signals from both CNN and CaiT models, using their probability estimates. This strategy is designed to identify strong agreement between the models while filtering out less confident signals. At each portfolio creation date, the predicted probabilities from both CNN and CaiT models are obtained. Each model independently ranks all stocks, sorting them into deciles based on their probability estimates. After this sorting, stocks that fall into the highest decile in both models, indicating that both CNN and CaiT assign them among the highest probabilities for that date, are selected for long positions. Similarly, stocks that fall into the lowest decile in both models, meaning that both models classify them as among the least likely to perform well, are selected for short positions. All selected stocks are equally weighted within the portfolio.

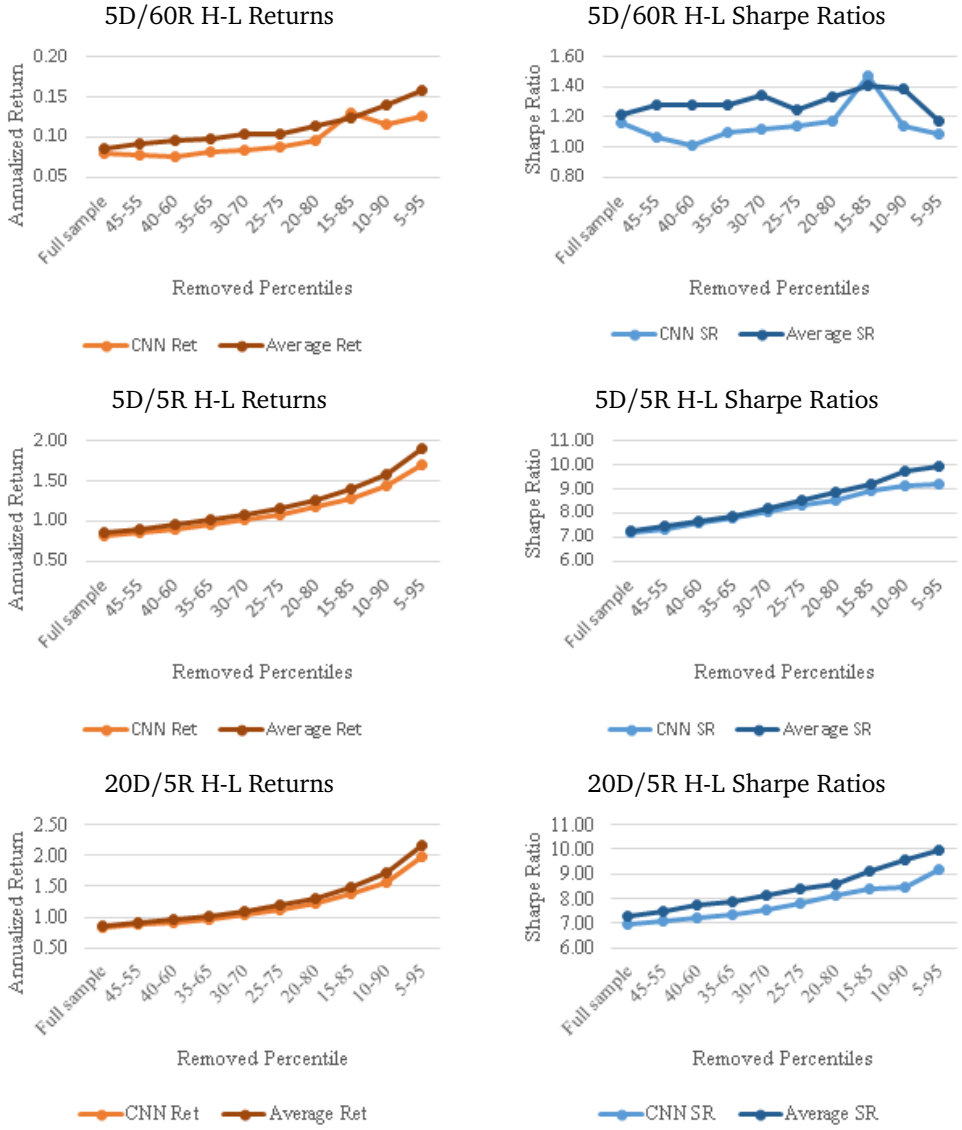

Figure 6: Effect of Confidence Filtering on H-L Returns and Sharpe Ratios

Description. The figure shows how H-L returns and Sharpe ratios change as low-confidence CNN probabilities are filtered. The top, middle, and bottom rows represent 5D/60R, 5D/5R, and 20D/5R, respectively. The CNN model is compared to the Average model, which applies filtering based on CNN probabilities. The x-axis shows removed probability percentiles, with "Full Sample" meaning all stocks are included. Percentile ranges indicate the stocks within those ranges are filtered out.

This approach offers two key advantages. First, it ensures that both models strongly agree on the highest and lowest deciles, making it possible to effectively combine their predictive strengths. At each rebalancing date, CNN captures local features while CaiT identifies global patterns, and only stocks that both models consistently classify as having the highest probability to rise or fall are included in the portfolio. This creates a stronger agreement than simple probability averaging, as it directly aligns the models' most confident predictions.

Second, this strategy acts as an implicit confidence filter. While it does not impose an absolute probability threshold, it inherently removes weaker signals by requiring that stocks be ranked in the top or bottom 10% of probabilities within both models. This ensures that even if a signal is not exceptionally strong, it must still be among the most confident signals relative to the full stock universe. Additionally, since probabilities are ranked within each model, label smoothing does not distort rankings, as a confident signal remains highly ranked even if its absolute probability estimate is slightly adjusted.

This investment strategy aligns with my empirical results, which demonstrate that removing weaker probability signals improves performance, and that combining CNN and CaiT probabilities leads to superior portfolio returns and Sharpe ratios. By ensuring that only high-confidence, mutually agreed-upon stocks are included, this strategy leverages the

complementary strengths of both models, leading to a more robust and reliable stock selection process.

6.2 Investment Performance

In this section, I analyze the performance of the proposed investment strategy by evaluating its gross returns, the impact of transaction costs, and cumulative performance over time. The strategy leverages both CNN and CaiT models to identify high-confidence long and short positions, ensuring that only stocks with strong agreement between the models are included. The results demonstrate that this approach consistently outperforms CNN portfolios and remains robust even after accounting for transaction costs, while its cumulative performance trajectory shows sustained long-term profitability.

Table 8 presents the raw performance of the investment strategy across different configurations (5D/5R, 20D/5R, and 5D/60R). The strategy delivers significant positive returns and high Sharpe ratios across all horizons. The strategy portfolio returns are 1.05 for 5D/5R, 1.14 for 20D/5R, and 0.11 for 5D/60R, with corresponding Sharpe ratios of 8.00, 8.29, and 1.33, respectively. These figures indicate that the strategy consistently generates strong risk-adjusted returns, with shorter-horizon models (5D/5R and 20D/5R) delivering the highest performance. Compared to CNN portfolios, these results show a clear improvement in both return levels and Sharpe ratios, highlighting the advantage of integrating CaiT's global feature extraction with CNN's local pattern recognition.

Table 9 assesses the strategy's performance after accounting for transaction costs. Despite the high turnover rates of 708% for 5D/5R and 700% for 20D/5R, the strategy remains highly profitable. The strategy portfolio returns after transaction costs are 0.70 for 5D/5R, 0.80 for 20D/5R, and 0.08 for 5D/60R, with Sharpe ratios of 5.37, 5.82, and 1.04, respectively. While transaction costs reduce returns, the strategy still outperforms CNN and maintains strong positive Sharpe ratios, demonstrating its resilience to trading frictions.

Figure 7 provides further validation by displaying the cumulative volatility-adjusted log returns of the strategy compared to various benchmarks, including CNN H-L portfolios, Momentum, Short-term reversal and Weekly Short-term Reversal strategies, and the S&P 500 (SPY). The proposed strategy exhibits the highest cumulative return trajectory, consistently outperforming both CNN models and other trend-based and market-based benchmarks. This sustained outperformance reinforces the effectiveness of combining CNN and CaiT models in a structured investment framework.

Overall, these results confirm that the proposed strategy delivers superior returns and risk-adjusted performance compared to CNN models. Even after incorporating transaction costs, the investment strategy remains highly profitable, highlighting its practical feasibility in real-world trading environments. The combination of deep learning-based stock selection and confidence-based filtering enhances portfolio efficiency, making it a robust and effective approach to quantitative investing.

7 Discussion and Limitations

In this section, I will discuss my findings and acknowledge the study's limitations. This paper aims to achieve three main objectives. The first objective is to replicate the study by (Jiang et al., 2023) to evaluate whether CNNs can autonomously extract meaningful features from stock chart images and whether these features possess predictive power. The second objective is to assess whether implementing a Vision Transformer architecture, specifically CaiT, using the same data and structure as the CNN model, can produce superior results. The final objective is to determine whether combining CNN and Vision Transformer predictions in an investment strategy can enhance the performance beyond that of standalone CNN or CaiT models. The discussion will be structured around these three key points.

Jiang et al. (2023) find that CNN-based strategies exhibit strong predictive power, particularly over short-term horizons, such as on a weekly basis. Their results indicate that the lowest decile portfolios tend to have substantially negative Sharpe ratios, while the highest decile portfolios demonstrate significantly positive values. Furthermore, they observe a monotonic increase in Sharpe ratios from the lowest to the highest decile Jiang et al. (2023, p. 3207).

My findings corroborate these results. In weekly strategies, CNN-based models exhibit Sharpe ratios of approximately -2.0 in the lowest decile, while the highest decile exceeds 2.7. Additionally, I observe a similar monotonic increase across deciles. Jiang et al. (2023) also find that image-based strategies generate highly effective trading signals, outperforming traditional price-based signals, albeit at the cost of significantly higher turnover Jiang et al. (2023, p. 3210). My results confirm that weekly CNN-based models achieve superior portfolio performance, with the highest H-L Sharpe ratio for equal-weighted portfolios reaching 7.2, compared to 2.91 for the best-performing traditional trend-following strategy (WSTR). However, these high returns are accompanied by considerable portfolio turnover, which is slightly higher than that of WSTR, nearly four times greater than MOM, and twice that of the STR.

To determine whether the outperformance of CNN-based strategies is primarily driven by high turnover, (Jiang et al., 2023) extend their analysis to longer trading horizons, examining monthly and quarterly portfolios. Their results indicate that CNN-based strategies continue to outperform traditional signals even at reduced turnover levels. Furthermore, they find that CNN-based models achieve higher Sharpe ratios than momentum strategies with comparable turnover levels Jiang et al. (2023, p. 3211). My findings align with these conclusions. Monthly and quarterly CNN-based strategies—particularly 5D/20R, 20D/20R, and 5D/60R—consistently outperform traditional approaches over extended investment horizons. Additionally, quarterly CNN strategies maintain

[6] Modified, taken from Jiang et al. (2023, p. 3208)

Table 8: Investment Strategy Performance

	5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR
Short	-0.40	-2.86	-0.50	-3.09	0.04	0.18
Long	0.65	3.42	0.64	3.62	0.15	0.68
Strategy	1.05	8.00	1.14	8.29	0.11	1.33
Turnover	708%		700%		63%	

Description. The table presents the performance metrics of the investment strategy across three configurations: 5D/5R, 20D/5R, and 5D/60R. It reports returns (ret) and Sharpe ratios (SR) for short positions, long positions, and the strategy, which is formed as a long-short portfolio. Additionally, turnover rates are provided for each configuration, indicating the level of trading activity within the strategy.

Table 9: Investment Strategy Performance After Transaction Costs

	5D/5R		20D/5R		5D/60R	
	ret	SR	ret	SR	ret	SR
Short	-0.21	-1.55	-0.31	-1.93	0.06	0.23
Long	0.48	2.55	0.49	2.74	0.14	0.64
Strategy	0.70	5.37	0.80	5.82	0.08	1.04
Turnover	708%		700%		63%	

Description. The table presents the performance metrics of the investment strategy after accounting for transaction costs across three configurations: 5D/5R, 20D/5R, and 5D/60R. It reports returns (ret) and Sharpe ratios (SR) for short positions, long positions, and the strategy, which is structured as a long-short portfolio. Additionally, turnover rates are provided for each configuration, reflecting the level of trading activity after incorporating transaction costs.

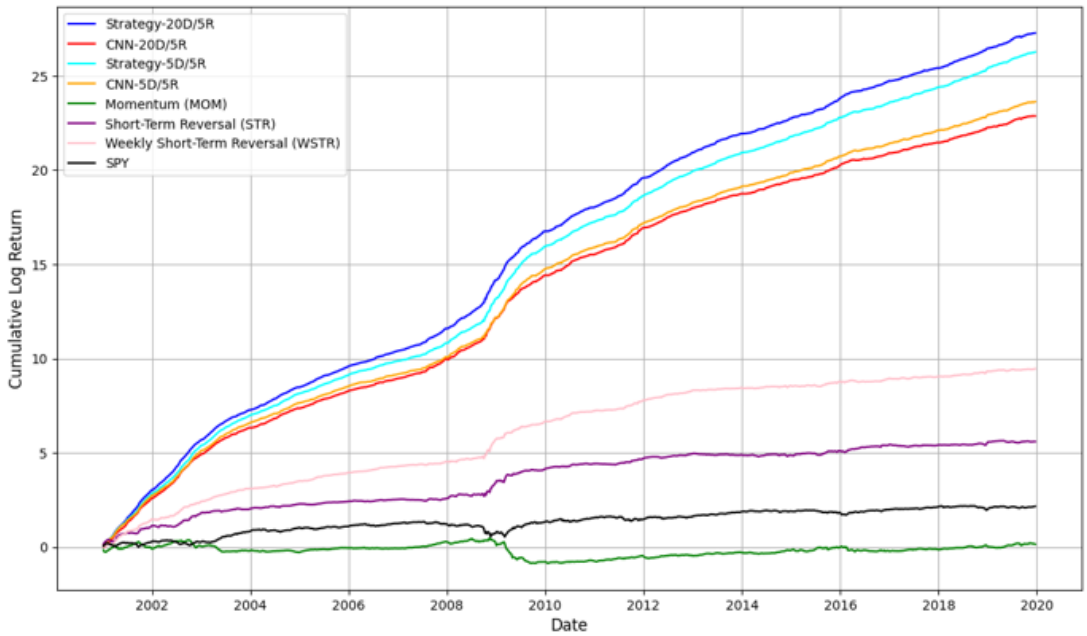

Figure 7: Cumulative Volatility Adjusted Returns of Equal-Weighted Portfolios[6]

Description. The figure presents the cumulative volatility-adjusted log returns of equal-weighted H-L portfolios for Strategy-20D/5R, CNN-20D/5R, Strategy-5D/5R, CNN-5D/5R, Momentum (MOM), Short-Term Reversal (STR), and Weekly Short-Term Reversal (WSTR). The SPY index is included for reference. All strategies have been adjusted to maintain the same volatility level as SPY over the test period.

turnover rates similar to those of monthly momentum strategies while still delivering superior Sharpe ratios for the H-L portfolio. Moreover, (Jiang et al., 2023) analyze transaction costs and cumulative log returns, concluding that CNN-based strategies remain profitable after accounting for transaction costs and continue to outperform traditional signals over time. My findings support this conclusion, as CNN-based strategies maintain their effectiveness even after transaction

cost adjustments and exhibit sustained cumulative outperformance compared to traditional approaches, as shown in Table 5 and Figure 7.

It is important to note that the primary objective of (Jiang et al., 2023) is not to develop a highly accurate return prediction model but rather to assess the predictive capabilities of image-based machine learning models in financial markets Jiang et al. (2023, p. 3204). These findings reinforce the conclusion that CNN-based strategies exhibit strong predictive power across both short-term and long-term horizons.

Now, I turn to the CaiT perspective to evaluate whether Vision Transformers enhance predictive performance compared to CNNs. The motivation for implementing a Vision Transformer architecture stems from its ability to capture global context, whereas CNNs are more effective at extracting local features. Given this distinction, CaiT may be better suited for identifying complex relationships in larger image contexts. To test this hypothesis, I specifically examine 5D/5R and 20D/5R configurations, as both models analyze the same stocks on the same dates but differ in the length of historical data included in their images. This setup results in four distinct models—CNN-5D/5R, CNN-20D/5R, CaiT-5D/5R, and CaiT-20D/5R—all making predictions on the exact same stocks. Therefore, I will compare their respective strengths in processing the same stocks. In addition to discussing returns and Sharpe ratios, I will also mention standard deviations to provide a more comprehensive assessment of risk. Furthermore, my discussion will focus exclusively on equal-weighted portfolios.

In Table 4, in the Low decile, CNN-20D/5R predicts the lowest returns and exhibits the lowest Sharpe ratio of -2.06, whereas CaiT-20D/5R achieves the highest Sharpe ratio among the models at -1.69. This suggests that, for the same set of stocks, CNN provides the most effective predictions in the lowest decile, when utilizing 20-day images.

For the High decile, CNN-5D/5R delivers the highest returns along with the highest Sharpe ratio, indicating that CNN remains the most effective predictor for stocks with strong upward momentum. However, an analysis of standard deviations in the High decile reveals that CaiT-20D/5R exhibits the lowest volatility among the four models. This suggests that while CaiT does not necessarily yield higher returns, it may be more effective at capturing global context, leading to more stable predictions with reduced variability. This ability to categorize stocks that move together more effectively could be one of CaiT's key strengths.

When evaluating Sharpe ratios of H-L portfolios, CNN-5D/5R and CaiT-20D/5R emerge as the top-performing models. CNN-5D/5R excels in short-term assessments by effectively capturing local patterns, resulting in the highest returns. In contrast, CaiT-20D/5R demonstrates an ability to group stocks with similar movement patterns more effectively in larger images, contributing to a more stable portfolio composition. This highlights the complementary nature of the two architectures—CNN excels in short-term feature extraction, while CaiT enhances global pattern recognition over extended time frames.

Recognizing that CNN and CaiT extract complementary information from stock chart images—with CNN focusing on predictive patterns that drive returns and CaiT emphasizing volatility reduction and stability—leads to a critical question: Can these two architectures be effectively combined to create an investment strategy that leverages their respective strengths? A key insight comes from simple probability averaging. In Table 7, the Average-20D/5R model achieved a higher H-L Sharpe ratio (7.31) compared to Average-5D/5R. This improved Sharpe ratio appears to stem from two distinct effects: a lower Sharpe ratio in the Low decile, likely attributed to CNN's ability to identify underperforming stocks, and a lower standard deviation in the High decile, potentially driven by CaiT's strength in capturing global patterns and reducing volatility. These observations suggest that combining CNN and CaiT predictions may create a more balanced and effective investment strategy, leveraging both architectures to optimize return potential while maintaining stability.

Another important aspect to consider is the role of confident signals. As demonstrated in Figure 6, filtering out less confident signals results in higher returns and Sharpe ratios. This effect may be attributed to two possible explanations. First, both CNN and CaiT architectures may struggle to accurately assess certain images, indicating a potential limitation in their ability to extract meaningful patterns from all stock charts. This could suggest that some price movements are inherently more difficult to predict using these models. Alternatively, a more speculative but intriguing possibility is that not all stock charts contain recognizable predictive patterns. Traditional technical analysis assumes that price patterns provide signals about future movements. However, the results may suggest that such patterns only appear in a specific subset of images, while others lack predictive features altogether. If this is the case, filtering out less confident predictions could serve as a way to identify and isolate the most informative patterns, improving overall strategy performance.

By combining both architectures, I develop an investment strategy that selects stocks classified in the High and Low deciles by both models. In Table 8, for the 20D/5R strategy, the portfolio exhibits higher returns on the Short leg and significantly lower standard deviation on the Long leg. This mirrors the patterns observed in Section 5.2 when evaluating individual models, in Section 5.4 when averaging their predictions. However, a key distinction arises in the investment strategy. The improved performance is not driven solely by one architecture but rather by the agreement between both models. Unlike previous observations, where I suggested that the effect might be attributed to a single model, in this case, both models independently select the same stocks. This alignment, rather than the influence of any individual model, is what ultimately drives the higher performance.

Yet, this also highlights certain limitations of the strategy. One key issue is that the agreement between models is purely mechanical—it is not integrated into the training process or loss function. A more advanced approach would involve designing an architecture that combines CNNs and Vision Transformers, allowing both models to influence the

learning process and jointly optimize parameters. Another limitation is that this strategy is inherently relative rather than absolute. It does not evaluate individual stocks in isolation but instead ranks them against each other. As a result, the strategy is ineffective in scenarios where the number of investable stocks is extremely limited—for instance, if only one stock were available, the strategy would not function. Finally, this study focuses on 5-day and 20-day images for CaiT, but 60-day images could provide new insights as the model may benefit from capturing longer-term dependencies in stock movements.

8 Conclusion

This study explores the predictive power of deep learning models applied to stock chart images, comparing Convolutional Neural Networks with Vision Transformers and analyzing their integration into an investment strategy. The research follows three main objectives: first, replicating (Jiang et al., 2023) to evaluate whether CNNs can autonomously extract predictive features from stock images; second, assessing whether CaiT, as a Vision Transformer model, can improve predictive performance by capturing global dependencies; and third, investigating whether combining CNN and CaiT predictions can enhance portfolio performance beyond standalone models. Using a dataset of stock images with 5-day, 20-day, and 60-day lookback periods, I trained multiple CNN and CaiT architectures, evaluated their returns, and Sharpe ratios, and examined their performance at different investment horizons.

For CNN-based strategies, my findings confirm those of (Jiang et al., 2023). CNN models effectively extract local features from stock images, generating profitable trading signals and strong Sharpe ratios, particularly in short-term horizons, while still maintaining predictive power in longer-term horizons, albeit to a lesser extent.

The study also examines CaiT's predictive capability, leveraging its strength in capturing global dependencies within stock chart images. While CNN excels at identifying high-return stocks, CaiT enhances portfolio stability by reducing volatility. Both models achieve the highest Sharpe ratios, but through different mechanisms: CNN focuses on shorter images, identifying stocks with strong return potential, whereas CaiT processes longer images, minimizing standard deviation. This suggests that CaiT may be more effective at identifying stocks that move together, contributing to lower portfolio volatility. These findings indicate that CNN and CaiT extract complementary features, supporting the hypothesis that combining both models could enhance predictive performance.

Building on these insights, I develop an investment strategy that integrates CNN and CaiT predictions by selecting stocks classified in the High and Low deciles by both models. This agreement-based approach effectively filters out weaker signals, leading to higher Sharpe ratios compared to using CNN or CaiT alone. Even after accounting for transaction costs, the strategy remains highly profitable, reinforcing its practical viability in real-world trading applications.

However, this strategy has limitations. The agreement between models is rule-based rather than learned during training. Additionally, the strategy is relative rather than absolute, meaning that it ranks stocks rather than predicting their absolute returns. This structure may limit its applicability in markets with fewer investable assets.

Overall, this research highlights the effectiveness of deep learning in financial prediction and demonstrates that CNN and CaiT offer strong predictive power focusing on complementary things. Future research could explore more sophisticated architectures that integrate both models into a unified framework, potentially improving performance beyond the agreement-based approach used here.

References

Allen, F., & Karjalainen, R. (1999). Using Genetic Algorithms to Find Technical Trading Rules. *Journal of Financial Economics*, *51*(2), 245–271. https://doi.org/10.1016/S0304-405X(98)00052-X

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. https://doi.org/10.48550/arXiv.1607.06450

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., et al. (2018). Relational Inductive Biases, Deep Learning, and Graph Networks. https://doi.org/10.48550/arXiv.1806.01261

Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance*, *47*, 1731–1764.

Brown, M. S., Pelosi, M. J., & Dirska, H. (2013). Dynamic-Radius Species-Conserving Genetic Algorithm for the Financial Forecasting of Dow Jones Index Stocks. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 27–41, Vol. 7988). Springer. https://doi.org/10.1007/978-3-642-39712-7_3

Cao, L. J., & Tay, F. H. (2003). Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting. *IEEE Transactions on Neural Networks*, *14*(6), 1506–1518. https://doi.org/10.1109/TNN.2003.820556

Chen, J.-F., Chen, W.-L., Huang, C.-P., Huang, S.-H., & Chen, A.-P. (2016). Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks, 87–92. https://doi.org/10.1109/CCBD.2016.027

Cohen, N., Balch, T., & Veloso, M. (2020). Trading via Image Classification. https://doi.org/10.48550/arXiv.1907.10046

Courant, R., Edberg, M., Dufour, N., & Kalogeiton, V. (2023). Transformers and Visual Transformers. In O. Colliot (Ed.), *Machine Learning for Brain* (pp. 193–233). Humana Press.

Cowles, A. (1933). Can Stock Market Forecasters Forecast? *Econometrica*, *1*(3), 309. https://doi.org/10.2307/1907042

Darji, M. (2021). Implementing Bilinear Interpolation for Image Resizing. *Medium*. https://meghal-darji.medium.com/implementing-bilinear-interpolation-for-image-resizing-357cbb2c2722

Detzel, A., Liu, H., Strauss, J., Zhou, G., & Zhu, Y. (2021). Learning and Predictability via Technical Analysis: Evidence from Bitcoin and Stocks with Hard-to-Value Fundamentals. *Financial Management*, *50*(1), 107–137. https://doi.org/10.1111/fima.12310

Di Persio, L., & Honchar, O. (2016). Artificial Neural Networks Approach to the Forecast of Stock Market Price Movements. *International Journal of Economics and Management Systems*, *1*, 158–162.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Fama, E. F., & Blume, M. E. (1966). Filter Rules and Stock-Market Trading. *The Journal of Business*, *39*(1), 226–241.

Frazzini, A., Israel, R., & Moskowitz, T. J. (2018). Trading Costs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3229719

Gani, H., Naseer, M., & Yaqub, M. (2022). How to Train Vision Transformer on Small-Scale Datasets? https://doi.org/10.48550/arXiv.2210.07240

Glorot, X., & Bengio, Y. (2010). Understanding the Difficulty of Training Deep Feedforward Neural Networks.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday Prediction of Borsa Istanbul Using Convolutional Neural Networks and Feature Correlations. *Knowledge-Based Systems*, *137*, 138–148. https://doi.org/10.1016/j.knosys.2017.09.023

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using Artificial Neural Network Models in Stock Market Index Prediction. *Expert Systems with Applications*, *38*(8), 10389–10397. https://doi.org/10.1016/j.eswa.2011.02.068

Han, Y., Liu, Y., Zhou, G., & Zhu, Y. (2021). Technical Analysis in the Stock Market: A Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3850494

Han, Y., Zhou, G., & Zhu, Y. (2016). A Trend Factor: Any Economic Gains from Using Information over Investment Horizons? *Journal of Financial Economics*, *122*(2), 352–375. https://doi.org/10.1016/j.jfineco.2016.01.029

Haussler, D. (1988). Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. *Artificial Intelligence*, *36*(2), 177–221. https://doi.org/10.1016/0004-3702(88)90002-1

He, K., Zhang, X. [, Ren, S., & Sun, J. (2016a). Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

He, K., Zhang, X. [, Ren, S., & Sun, J. (2016b). Identity Mappings in Deep Residual Networks. https://doi.org/10.48550/arXiv.1603.05027

Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). https://doi.org/10.48550/arXiv.1606.08415

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., & Oh, S. J. (2021). Rethinking Spatial Dimensions of Vision Transformers. https://doi.org/10.48550/arXiv.2103.16302

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. The MIT Press.

Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-Based Stock Market Prediction Using a Diverse Set of Variables. *Expert Systems with Applications*, *129*, 273–285. https://doi.org/10.1016/j.eswa.2019.03.029

Hu, G., Hu, Y. [, Yang, K., Yu, Z., Sung, F., Zhang, Z., Xie, F., Liu, J., Robertson, N., Hospedales, T., & Miemie, Q. (2018). Deep Stock Representation Learning: From Candlestick Charts to Investment Decisions, 2706–2710. https://doi.org/10.1109/ICASSP.2018.8462215

Hu, Y., Feng, B., Zhang, X. [, Ngai, E., & Liu, M. (2015). Stock Trading Rule Discovery with an Evolutionary Trend Following Model. *Expert Systems with Applications*, *42*(1), 212–222. https://doi.org/10.1016/j.eswa.2014.07.059

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep Networks with Stochastic Depth. *9908*, 646–661. https://doi.org/10.1007/978-3-319-46493-0_39

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/arXiv.1502.03167

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Jegadeesh, N., & Titman, S. (2001). Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *The Journal of Finance*, *56*(2), 699–720.

Jensen, M. C., & Benington, G. A. (1970). Random Walks and Technical Theories: Some Additional Evidence. *The Journal of Finance*, *25*(2), 469–482. https://doi.org/10.2307/2325495

Jiang, J., Kelly, B., & Xiu, D. (2023). (Re-)Imag(in)ing Price Trends. *The Journal of Finance*, *78*(6), 3193–3249. https://doi.org/10.1111/jofi.13268

Kamath, U., Graham, K., & Emara, W. (2022). *Transformers for Machine Learning: A Deep Dive (1st ed.)* Chapman & Hall.

Ke, Z., Kelly, B. T., & Xiu, D. (2021). Predicting Returns with Text Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3389884

Kercheval, A. N., & Zhang, Y. (2015). Modelling High-Frequency Limit Order Book Dynamics with Support Vector Machines. *Quantitative Finance*, *15*(8), 1315–1329. https://doi.org/10.1080/14697688.2015.1032546

Kim, T., & Kim, H. Y. (2019). Forecasting Stock Prices with a Feature Fusion LSTM-CNN Model Using Different Representations of the Same Data. *PLOS ONE*, *14*(2), e0212320. https://doi.org/10.1371/journal.pone.0212320

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://doi.org/10.1145/3065386

Lecun, Y., & Bengio, Y. (1995). Convolutional Networks for Images, Speech, and Time-Series. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*. The MIT Press.

Lee, J., Kim, R., Koh, Y., & Kang, J. (2019). Global Stock Market Prediction Based on Stock Chart Images Using Deep Q-Network. *IEEE Access*, *7*, 167260–167277. https://doi.org/10.1109/ACCESS.2019.2953542

Lee, M. (2009). Using Support Vector Machine with a Hybrid Feature Selection Method for Stock Trend Prediction. *Expert Systems with Applications*, *36*(8), 10896–10904. https://doi.org/10.1016/j.eswa.2009.02.038

Lee, S. H., Lee, S., & Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*. https://doi.org/10.48550/arXiv.2112.13492

Li, Y., Wang, J. [, Dai, X., Wang, L., Yeh, C. C. M., Zheng, Y., Zhang, W., & Ma, K. L. (2023). How Does Attention Work in Vision Transformers? A Visual Analytics Attempt. https://doi.org/10.48550/arXiv.2303.13731

Lin, T., Wang, Y. [, Liu, X., & Qiu, X. (2022). A Survey of Transformers. *AI Open*, *3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Lo, A. W., & Hasanhodzic, J. (2010). *The Heretics of Finance: Conversations with Leading Practitioners of Technical Analysis* (Vol. 16). John Wiley and Sons.

Lo, A. W., Mamaysky, H., & Wang, J. [(2000). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance*, *55*(4), 1705–1765. https://doi.org/10.1111/0022-1082.00265

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. https://doi.org/10.48550/arXiv.1711.05101

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. *30th International Conference on Machine Learning*.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9), Article 5521. https://doi.org/10.3390/app13095521

Menkhoff, L., & Taylor, M. P. (2007). The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. *Journal of Economic Literature*, *45*(4), 936–972.

Murray, S., Xia, Y., & Xiao, H. (2024). Charting by Machines. *Journal of Financial Economics*, *153*, Article 103791. https://doi.org/10.1016/j.jfineco.2024.103791

Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science*, *60*(7), 1772–1791.

Nelson, D. M. Q., Pereira, A. C. M., & Oliveira, R. A. d. (2017). Stock Market's Price Movement Prediction with LSTM Neural Networks. *2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA*.

Noble, W. S. (2006). What Is a Support Vector Machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Peng, Z., Guo, Z., Huang, W., Wang, Y. [, Xie, L., Jiao, J., Tian, Q., & Ye, Q. (2023). Conformer: Local Features Coupling Global Representations for Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 9454–9468. https://doi.org/10.1109/TPAMI.2023.3243048

Raschka, S. (2024). *Build a Large Language Model (From Scratch)*. Manning Publications.

Shao, R., & Bi, X.-J. (2022). Transformers Meet Small Datasets. *IEEE Access*, *10*, 118454–118464. https://doi.org/10.1109/ACCESS.2022.3221138

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, *15*(56), 1929–1958.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Finance*, *54*(5), 1647–1691.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. https://doi.org/10.48550/arXiv.1409.4842

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training Data-Efficient Image Transformers & Distillation Through Attention. https://doi.org/10.48550/arXiv.2012.12877

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going Deeper with Image Transformers. https://doi.org/10.48550/arXiv.2103.17239

Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., & Wong, T.-L. (2007). Design and Implementation of NN5 for Hong Kong Stock Price Forecasting. *Engineering Applications of Artificial Intelligence*, *20*(4), 453–461. https://doi.org/10.1016/j.engappai.2006.10.002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Zhu, Y., & Zhou, G. (2009). Technical Analysis: An Asset Allocation Perspective on the Use of Moving Averages. *Journal of Financial Economics*, *92*(3), 519–544. https://doi.org/10.1016/j.jfineco.2008.07.002

List of Abbreviations

CaiT Class-Attention in Image Transformers

CNN Convolutional Neural Network

CRSP Center for Research in Security Prices

DeiT Data-efficient Image Transformers

EMH Efficient Market Hypothesis

FNN Feedforward Neural Network

GA Genetic Algorithm

LSA Locality Self-Attention

LSTM Long Short-Term Memory

MA Moving Average

ML Machine Learning

MLP Multi-Layer Perceptron

MOM Momentum

NLP Natural Language Processing

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

RSI Relative Strength Index

SPT Shifted Patch Tokenization

STR Short-term Reversal

SVM Support Vector Machine

ViT Vision Transformer

WSTR Weekly Short-term Reversal

Junior Management Science 11(1) (2026) 227-245

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Relative Performance-Messung als Instrument der Unternehmenssteuerung

Felix Achim Bautz

Heinrich Heine University Düsseldorf

Abstract

Stabilizing and improving organizational performance is a central challenge in management control. Behavior-based results controls is a key mechanism in this context. The lack of a generally effective and accepted incentive system requires a deeper examination of the underlying mechanisms beyond surface-level system features. Relative Performance Measurement (RPM) is considered a dynamic tool to align performance and motivation with strategic goals. However, its use is associated with risks such as demotivation and dysfunctional behavior. This study analyzes the effects of RPM in the context of competition and social comparison within the social system of the organization. Based on agency theory (AT), the economic rationale of RPM is derived using tournament theory. Its limitations are discussed using the example of forced-ranking practices. The perspective is extended by social comparison theory, which shows how task-related, individual, and contextual variables, as well as system design, influence the effectiveness of RPM in different settings. The study contributes to the integration of economic and behavioral approaches in performance measurement (PM) and provides practical implications for the targeted design of RPM systems in management control

Zusammenfassung

Die Stabilisierung und Steigerung der Unternehmensleistung stellt eine zentrale Herausforderung der Unternehmenssteuerung dar. Dabei kommt der verhaltensorientierten Ergebnissteuerung eine besondere Bedeutung zu. Das Fehlen eines allgemein akzeptierten und wirksamen Anreizsystems erfordert eine vertiefte Auseinandersetzung mit den zugrundeliegenden Wirkmechanismen jenseits oberflächlicher Systemmerkmale. Relative Performance-Messung (RPM) bietet in diesem Kontext ein dynamisches Instrument zur strategiekonformen Ausrichtung von Leistung und Motivation, ist jedoch zugleich mit Risiken wie Demotivation und dysfunktionalem Verhalten verbunden. Diese Studie analysiert die Wirkung von RPM im Spannungsfeld von Wettbewerb und sozialem Vergleich, eingebettet in das soziale System „Unternehmen“. Aufbauend auf der Agency-Theorie (AT) wird die ökonomische Intuition von RPM durch die Turniertheorie hergeleitet, deren Grenzen durch die Analyse der Forced-Ranking-Praxis aufgezeigt werden. Die Perspektive wird durch die Theorie des sozialen Vergleichs (TSV) erweitert, wodurch aufgezeigt werden kann, wie aufgabenbezogene, personenbezogene und umgebungsbezogene Variablen sowie das Design des Anreizsystems die Wirksamkeit von RPM in unterschiedlichen Anwendungsszenarien beeinflussen. Die Arbeit leistet damit einen Beitrag zur Integration ökonomischer und sozial-psychologischer Ansätze der Performance-Messung (PM) und formuliert praxisorientierte Implikationen für die passgenaue Gestaltung von RPM-Systemen in der Unternehmenssteuerung.

Keywords: turnier-theorie; theorie des sozialen vergleichs; unternehmenssteuerung; agency-theory; relative performance-messung

DOI: https://doi.org/10.5282/jums/v11i1pp227-245

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

1 Einleitung

Unternehmen investieren weltweit erhebliche Ressourcen in die Entwicklung und Implementierung von Systemen zur Unternehmenssteuerung, um das Verhalten ihrer Mitarbeitenden zu lenken und strategische Ziele zu erreichen (vgl. K. R. Murphy 2020, S. 22). Trotz langjähriger Forschung und zahlreicher Optimierungsversuche bleibt die Unzufriedenheit nach wie vor hoch (vgl. S. Adler et al. 2016, S. 219; K. Murphy und A. DeNisi 2023, S. 151). Umfragen zeigen, dass stets eine Mehrheit der Unternehmen angeben, Systeme der Performance-Messung (PM) zu nutzen, um strategische Entscheidungen wie Vergütung und Beförderung zu steuern (vgl. K. Murphy und A. DeNisi 2023, S. 145; K. R. Murphy 2020, S. 15f). Eine zentrale Aufgabe des Controllings besteht darin, den Managementprozess von Zielfindung, Planung und Steuerung zu begleiten (vgl. E. Weißenberger 2007, S. 35). Dazu zählt auch die Konzeption und Weiterentwicklung von PM-Instrumenten im Sinne der Unternehmenssteuerung. Im Object-of-Control-Framework (OoC) nach A. Merchant und Van der Stede 2017 wird Unternehmenssteuerung als Management-Control (MC) interpretiert. Die Autoren betrachten aus verhaltensorientierter Perspektive, wie das Handeln der Mitarbeitenden zur Erreichung strategischer Ziele beiträgt. Sie identifizieren dabei drei zentrale Steuerungsprobleme: mangelnde Führung (*lack of direction*), mangelnde Motivation (*lack of motivation*) und fehlende persönliche Fähigkeiten (*personal limitations*). Diesen Herausforderungen kann laut dem OoC-Framework durch den gezielten Einsatz von Management-Control-Systems (MCS) begegnet werden (vgl. A. Merchant und Van der Stede 2017, S. 11–18; D. Otley 1999, S. 364; E. Strauß und C. Zecher 2013, S. 244–248). PM wird als MCS definiert, das zunächst die Bewertung der Performance umfasst, um anschließend die gewonnenen Daten in einem geeigneten Anreizsystem bereitzustellen. Es stellt die Voraussetzung zur Steuerung eigenständig handelnder Mitarbeitender dar (vgl. S. Adler et al. 2016, S. 219; P. Baker 1992, S. 598; A. Merchant und Van der Stede 2017, S. 355–357; S. Stewart et al. 2010, S. 171). Die erste Entscheidung im Rahmen der PM-Gestaltung ist die Wahl der Klasse der Messung. Diese Arbeit legt den Fokus auf relative Performance-Messung (RPM) als Alternative zur absoluten Performance-Messung (APM) (vgl. E. Duffy und E. Webber 1974, S. 307; D. Goffin et al. 1996, S. 23f; H. Wruck und Y. Wu 2022, S. 2951). RPM verspricht theoretisch Schwächen von APM zu reduzieren, die direkten Kosten zu senken und bietet vor allem eigene Vorteile. Dennoch ist eine kritische Auseinandersetzung mit unbeabsichtigten Folgen und indirekten Kosten von RPM erforderlich (vgl. R. Frederickson 1992, S. 652; E. Goodhart 1984, S. 5; F. Ridgway 1956, S. 240f; D. Tafkov 2013, S. 329). RPM ist ein weit verbreitetes und regelmäßig eingesetztes Instrument in Unternehmen (vgl. S. Kramer et al. 2016, S. 16; H. Song et al. 2018, S. 2628). Die Mechanik von RPM basiert darauf, evolutionäre Antriebe wie Wettbewerb und sozialen Vergleich (SV) in monetäre Gewinne oder Anerkennung zu übersetzen (vgl. A. Tran und R. Zeckhauser 2012, S. 645). Die Literatur zeigt, dass RPM-Systeme auch ohne Vergütungskomponenten untersucht werden, um isolierte Verhaltenseffekte unabhängig von monetären Anreizen zu analysieren (vgl. L. Gerhards und N. Siemer 2016, S. 1189; L. Hannan et al. 2013, S. 554; C. Schnieder 2022, S. 73–75). Die Leitfrage lautet, vor welchen Herausforderungen die Gestaltung von Anreizsystemen auf Basis relativer Performance-Messung steht.

Die Agency-Theorie (AT) spielt eine zentrale Rolle in der Diskussion um die Wirksamkeit von RPM, da sie das grundlegende Problem der Informationsasymmetrie zwischen Prinzipal und Agenten adressiert (vgl. P. Baker 1992, S. 599f; T. Ball et al. 2020, S. 1507; S. Goebel und E. Weißenberger 2017, S. 191; J. Luft 2016, S. 75). Seit über vier Jahrzehnten belegen analytische Modelle die grundlegende Nützlichkeit von RPM, Anstrengung und Performance zu steigern. Diese Modelle zeigen in der Praxis nicht immer die prognostizierten Ergebnisse (vgl. T. Ball et al. 2020, S. 1508f; E. Dechenaux et al. 2015, S. 610–613; R. Frederickson 1992, S. 647–649; J. Luft 2016, S. 76–78; K. R. Murphy 2020, S. 13–15). Die Vielfalt in Branchen, Aufgabenbereichen und Persönlichkeiten erfordert, über rein ökonomische Grundlagen hinauszugehen und das Verständnis von Management-Buchhaltern für relevante Einflussfaktoren zu vertiefen (vgl. S. Goebel und E. Weißenberger 2017, S. 192). Die Forschung betrachtet RPM zunehmend durch die sozial-psychologische Perspektive, insb. die Theorie des sozialen Vergleichs (TSV). Dieser Ansatz berücksichtigt den menschlichen Grundmechanismus der Selbstbewertung und des Wettbewerbs in Fähigkeiten (vgl. L. Hannan et al. 2013, S. 554; S. Murthy und A. Schafer 2011, S. 159f.; C. Schnieder 2022, S. 72–75; H. Song et al. 2018, S. 2629; D. Tafkov 2013, S. 328). In Abwesenheit objektiver Standards treten Individuen in einen SV mit einer relevanten Vergleichsgruppe. Resultat sind eine aktualisierte Selbstdefinition und folglich eine affektive Verhaltensreaktion (vgl. G. Birnberg et al. 2006, S. 121–123; L. Festinger 1954, S. 117–123; A. Tesser 1988, S. 181–183). Eine fundierte Untersuchung sollte der Steuerung dieser Triebkraft in RPM-Systemen besondere Aufmerksamkeit widmen.

AT und TSV zeigen übereinstimmend auf, dass RPM in bestimmten Situationen nützlich sein kann. Eine wesentliche Voraussetzung ist eine ausreichende gemeinsame Unsicherheit unter den Agenten. Dies beschreibt ein kritisches Maß an Ähnlichkeit unter den Teilnehmenden, damit der SV-Prozess genügend Relevanz besitzt (vgl. R. Frederickson 1992, S. 653; D. Tafkov 2013, S. 330; A. Tesser 1988, S. 182). Weiterhin erfordert die Konzeption von RPM-Systemen eine ganzheitliche Betrachtung der Steuerungssituation. Neben den Effekten unterschiedlicher Darstellungsformen sind auch die Persönlichkeitsmerkmale der Mitarbeitenden sowie die Definition der Aufgaben zentral (vgl. E. Bonner und B. Sprinkle 2002, S. 305f.; M. Garcia et al. 2013, S. 635f.; L. Hannan et al. 2013, S. 554; C. Schnieder 2022, S. 75–77). Besonders sensibel reagieren leistungsstarke und leistungsschwache Mitarbeitende, deren unterschiedliche

Anforderungen an RPM-Systeme eine gezielte und differenzierte Gestaltung notwendig machen (vgl. H. Aguinis und K. J. Bradley 2015, S. 161f.; H. Eyring und G. Narayanan 2018, S. 681–584; S. Kramer et al. 2016, S. 17f.; K. R. Murphy 2020, S. 17–19; S. Murthy und A. Schafer 2011, S. 160f.). Kapitel zwei ordnet PM in das OoC-Framework ein, grenzt RPM von APM ab und liefert eine Definition von RPM. Darauf aufbauend wird der theoretische Rahmen im Kontext der Theorie angemessener Komplexität entwickelt. Kapitel drei untersucht die grundlegende ökonomische Nützlichkeit von RPM anhand der Turnier-Theorie. Ergänzend werden praxisrelevante Herausforderungen beleuchtet, um die Perspektive der Verhaltenstheorie einzubeziehen. Kapitel vier widmet sich der Konzeption eines Forschungsrahmens zur sozial-psychologisch geprägten Analyse von RPM und präsentiert wesentliche Forschungsergebnisse anhand dieser Kriterien. Im fünften Kapitel werden die gewonnenen Erkenntnisse zu Implikationen für eine effiziente Gestaltung von Anreizsystemen verdichtet. Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick auf zukünftige Forschungsansätze.

2 Anreizsysteme und relative Performance-Messung im Kontext der Unternehmenssteuerung

2.1 Einordnung der Performance-Messung in die Unternehmenssteuerung

Zentraler Fokus des Managements liegt in der Definition einer angestrebten Unternehmensleistung und Wege zu deren Erreichung. Eine erfolgreiche und zielkonforme Unternehmensleistung ist als Kombination aus einer guten Strategie und der effektiven Gestaltung der Steuerungssysteme zu interpretieren (vgl. S. Goebel und E. Weißenberger 2017, S. 189; A. Merchant und Van der Stede 2017, S. 11f.; K. R. Murphy 2020, S. 14; D. Otley 1999, S. 363f.). Der Managementprozess umfasst dabei im wesentlichen drei Aspekte. Zunächst werden übergeordnete richtungsweisende Ziele festgelegt. Nachfolgend kann aus dieser Grundlage heraus eine geeignete Strategie abgeleitet werden. Sie definiert langfristig ausgelegte Handlungspläne zur Nutzung von Ressourcen und der Zielerreichung (vgl. S. Goebel und E. Weißenberger 2017, S. 189; A. Merchant und Van der Stede 2017, S. 8f.; A. Merchant und T. Otley 2006, S. 785f.; D. Otley 1999, S. 363–365; E. Strauß und C. Zecher 2013, S. 244f.). Schließlich verfolgen MC und MCS laut OoC-Framework das Ziel, individuell handelnde Mitarbeitende entlang der Unternehmensziele zu steuern und das Erreichen der Strategie anzustreben. Also eine Kongruenz zwischen gewünschtem und tatsächlichem Verhalten herzustellen (vgl. A. Merchant und Van der Stede 2017, S. 11f.; A. Merchant und T. Otley 2006, S. 785f.; K. R. Murphy 2020, S. 14; E. Strauß und C. Zecher 2013, S. 244f.). A. Merchant und Van der Stede (2017, S. 10–12) stellen sich in diesem Zusammenhang die Frage, ob und von wem die Unternehmensziele verfolgt und erreicht werden. Aus dieser Frage heraus sind die drei Hauptprobleme im Verhalten zu identifizieren, denen MC durch Steuerungsaktivität begegnet. Erstens eine mangelnde Führung. Dabei wird eine Situation beschrieben, in der die Mitarbeitenden die Ziele und daraus entspringende Erwartungen des Unternehmens nicht verstehen, woraus eine gestörte Orientierung resultiert. Zweitens eine mangelnde Motivation. Die Mitarbeitenden verstehen zwar, was die Ziele sind, lassen ihr Handeln aber von Eigeninteresse dominieren und arbeiten nicht konsequent im Sinne oder sogar auf Kosten des Unternehmens. Zuletzt kann die Performancelücke auch im Fehlen relevanter Fähigkeiten begründet sein. Also in personenbezogenen Einschränkungen, die ein zielgerichtetes Verhalten verhindern (vgl. C. Kleine und E. Weißenberger 2014, S. 242; A. Merchant und Van der Stede 2017, S. 11–15; E. Strauß und C. Zecher 2013, S. 244f.; E. Weißenberger 2019, S. 154). Aus diesen Steuerungsproblemen heraus definieren die Autoren drei MCS im Rahmen des OoC-Frameworks: die Ergebnissteuerung, die Prozesssteuerung und die Mitarbeitersteuerung. Sie stellen den Werkzeugkasten des Managements dar, um sicherzustellen, dass alle an der Strategieumsetzung mitwirken (vgl. C. Kleine und E. Weißenberger 2014, S. 245; A. Merchant und Van der Stede 2017, S. 33–35, 86f., 97f.; K. R. Murphy 2020, S. 14; E. Strauß und C. Zecher 2013, S. 248f.; E. Weißenberger 2019, S. 154).

PM ist eine Form der Ergebnissteuerung. Zielobjekt dieser Steuerungssysteme ist die Performance der Mitarbeitenden, gemessen an ihren Ergebnisbeiträgen. Die Ergebnissteuerung regt zur Reflexion an, ohne Prozesse und Aktivitäten strikt vorzugeben. Sie fördert vielmehr die anreizgesteuerte Befähigung der Mitarbeitenden, Maßnahmen zu ergreifen, die den vom Unternehmen erwarteten Output erzielen und individuelle Interessen überwinden (vgl. C. Kleine und E. Weißenberger 2014, S. 244f.; A. Merchant und Van der Stede 2017, S. 33f., 335–337; E. Strauß und C. Zecher 2013, S. 246–248). PM-Systeme bestehen aus zwei Schritten. Einerseits aus der Evaluation der Performance und andererseits der Nutzung von gewonnenen Informationen zur Gestaltung eines Anreizsystems zur Verhaltenssteuerung (vgl. S. Adler et al. 2016, S. 219; P. Baker 1992, S. 598; A. Merchant und Van der Stede 2017, S. 355–357; S. Stewart et al. 2010, S. 171). Sie erzielen Motivationseffekte durch Belohnungen und Sanktionen. Belohnungen fördern gewünschtes Verhalten und umfassen finanzielle Anreize wie Boni, Gehaltserhöhungen oder Beförderungen sowie nicht-finanzielle Anreize wie Schulungen, Sachprämien oder Anerkennung in Form von Stolz und Reputation (vgl. A. Merchant und Van der Stede 2017, S. 33; D. Otley 1999, S. 368; C. Schnieder 2022, S. 73f.). Sanktionen hingegen verhindern unerwünschtes Verhalten durch Maßnahmen wie das Verwehren positiver Anreize, finanzielle Kürzungen oder öffentliche Kritik (vgl. N. Kluger und A. DeNisi 1996, S. 258f.; K. R. Murphy 2020, S. 14; S. Stewart et al. 2010, S. 169f.). Beide Mechanismen zielen darauf ab, das Verhalten der Mitarbeitenden auf die Unternehmensziele auszurichten (vgl. E. Bonner und B. Sprinkle 2002, S. 329; A. Merchant und Van der Stede 2017, S. 33–40; K. R. Murphy 2020, S. 14; C. Schnieder 2022, S. 73f.).

Entscheidet sich das Management zur Nutzung von PM als Anreizsystem im Zuge der Unternehmenssteuerung (Frage nach dem Ob), muss nachfolgend zwangsläufig überlegt werden, in welcher Art und Weise dies geschehen soll (Frage nach dem Wie). In erster Instanz stehen sich dabei die PM-Systemklassen APM und RPM gegenüber (vgl. E. Duffy und E. Webber 1974, S. 307; D. Goffin et al. 1996, S. 23f.; H. Wruck und Y. Wu 2022, S. 2951). APM-Systeme (Management by Objectives) richten die Steuerung von Mitarbeitenden an individuellen Leistungszielen aus, die aus der Unternehmensstrategie abgeleitet werden. Diese Ziele legen Standards auf Basis absoluter Kennzahlen fest, die zur Erlangung von Anreizen erreicht werden müssen (vgl. P. Baker 1992, S. 599; E. Duffy und E. Webber 1974, S. 307; D. Goffin et al. 1996, S. 24; A. Merchant und Van der Stede 2017, S. 38; D. Otley 1999, S. 368; H. Wruck und Y. Wu 2022, S. 2951). Im Gegensatz dazu vergleichen RPM-Systeme individuelle Beiträge innerhalb einer Performanceverteilung. Ein Standard bzw. die Bestperformance für die Messgröße wird durch die relative Positionierung gegenüber Kolleginnen und Kollegen in ähnlichen Rollen festgelegt (vgl. E. Duffy und E. Webber 1974, S. 23f.; H. Wruck und Y. Wu 2022, S. 2951). Gemäß der Kontingenztheorie existiert kein universell anwendbares PM-System, weshalb die Wahl eines passenden Systems für den spezifischen Anwendungsfall eine zentrale Herausforderung darstellt (vgl. D. Otley 1999, S. 244). Während APM-Systeme weit verbreitet sind, weisen sie strukturelle Schwächen auf. Darunter den Verlust der Steuerungsfunktion durch die einseitige Fokussierung auf Zielgrößen (Goodhart's Law), die zu dysfunktionalen Nebeneffekten führen können (vgl. P. Baker 1992, S. 599f.; E. Goodhart 1984, S. 5; A. Merchant und Van der Stede 2017, S. 33; F. Ridgway 1956, S. 240f.). RPM-Systeme bieten hier ein alternatives Potenzial, da sie auf andere Wirkungsmechanismen setzen und gleichzeitig eine vollständige Leistungsdifferenzierung aller Beteiligten ermöglichen (vgl. E. Duffy und E. Webber 1974, S. 307; R. Frederickson 1992, S. 652; D. Tafkov 2013, S. 329; A. Wijayanti et al. 2024, S. 2f.). Der Fokus dieser Arbeit liegt auf RPM, um dessen eigenständige, potenziell überlegene Nützlichkeit für die Unternehmenssteuerung zu untersuchen.

Die Definition von RPM geht über eine bloße Abgrenzung von APM hinaus und lässt sich nicht eindeutig festlegen. Vielmehr erfolgt eine präzisere Bestimmung durch die Beantwortung zweier konzeptioneller Fragestellungen. Einerseits stellt sich die Frage, auf wen sich der Vergleich bezieht. Im Kontext der Unternehmenssteuerung kann der Vergleich entweder auf die Managementebene fokussiert sein, wobei häufig ein unternehmensexterner Branchenbezug besteht. Betriebswirtschaftlich ebenso interessant ist die Steuerung von Angestellten aller weiteren Hierarchieebenen. Also der Vergleich gleicher bzw. ähnlicher Kollegen untereinander zur Performancesteigerung im unternehmensinternen Kontext (vgl. T. Ball et al. 2020, S. 1508; J. Luft 2016, S. 75f.; C. Schnieder 2022, S. 73–75). Andererseits muss entschieden werden, welche der drei in Abb. 1 dargestellten Vergleichsformen zur Anwendung kommt.

Ein RPM-System nutzt ordinalen Vergleich, wenn es eine

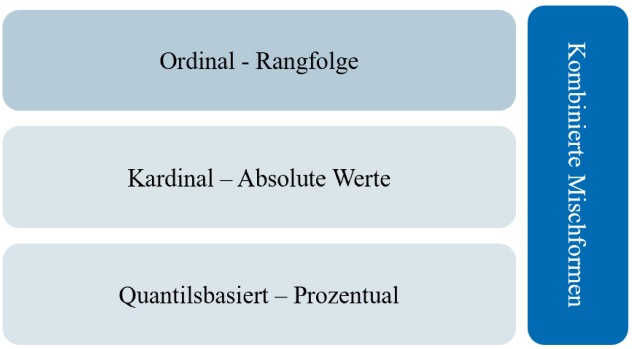

Abbildung 1: Formen der relativen Performance-Messung.

Rangfolge vom ersten bis zum n-ten Platz verwendet, ohne Informationen über die Performanceabstände zwischen den Rängen zu geben, z. B. „Ich bin der vierte von zehn in der Rangfolge" (vgl. E. Dechenaux et al. 2015, S. 610; D. Gill et al. 2019, S. 494; D. Goffin et al. 1996, S. 24; L. Hannan et al. 2008, S. 893f.; C. Schnieder 2022, S. 77f.; D. Tafkov 2013, S. 330f.). Wird im Performancevergleich ein RPM-System mit absoluten numerischen Werten gearbeitet, gilt ein kardinaler Vergleich, da hier die tatsächlichen Performancedifferenzen sichtbar werden (vgl. L. Hannan et al. 2019, S. 160; A. Mohnen et al. 2008, S. 693f.; C. Schnieder 2022, S. 77). Quantilsbasierte Systeme gruppieren die Performance in prozentuale Kategorien, die auf der relativen Position in der Verteilung basieren, oder ermöglichen den Vergleich mit einem prozentualen Referenzpunkt. Hierbei können verschiedene Kategorien und Genauigkeiten, wie Perzentile, Dezile oder Quartile, verwendet werden. Ein weiteres Beispiel ist Benchmarking, bei dem die individuelle Performance mit einem festgelegten Referenzwert verglichen wird, etwa der Frage, ob ein Mitarbeitender über oder unter dem Median der Leistungsverteilung liegt (vgl. Y. Chen et al. 2010, S. 1358f.; D. Goffin et al. 1996, S. 23; S. Murthy und A. Schafer 2011, S. 160f.; C. Schnieder 2022, S. 77). Kombinierte Mischformen sind ebenfalls möglich, bei denen ordinale Ränge mit kardinalen Werten verknüpft werden. Ebenso kann ein kardinaler Performancestandard mit einem Referenzpunkt kombiniert werden (vgl. C. Schnieder 2022, S. 78). Nach der allgemeinen Einordnung und Abgrenzung von RPM in diesem Kapitel wird im folgenden Abschnitt 2.2 die Wirksamkeit von RPM-Systemen näher analysiert.

2.2 Wirksamkeit der relativen Performance-Messung und ihre theoretischen Hintergründe

Die Situation, dass jahrzehntelange Forschung kein Ergebnis hervorgebracht hat, das allgemein als hilfreich akzeptiert ist, zeigt: Die Zeit der kosmetischen Anpassungen ist angezählt, und es bedarf einer grundlegenden Reform in der Konzeption von RPM-Systemen (vgl. S. Adler et al. 2016, S. 2109; K. Murphy und A. DeNisi 2023, S. 146; K. R. Murphy 2020, S. 13–15). K. Murphy und A. DeNisi (2023, S. 143–145) betonen, dass die Abkehr von PM

bzw. RPM keine Option ist. Die Ziele und Notwendigkeit laut OoC-Framework bleiben ungebrochen richtig und das Management würde sonst *„das Kind mit dem Badewasser ausschütten“* (K. Murphy und A. DeNisi 2023, S. 144) und ohne Kontrolle über das Verhalten der Mitarbeitenden navigieren (vgl. S. Adler et al. 2016, S. 219; K. Murphy und A. DeNisi 2023, S. 144f.). Vielmehr bedarf es einer ganzheitlichen Betrachtung des verhaltensorientierten Wirkungsmechanismus von RPM im Feld der wesentlichen Erklärungsansätze. Ziel muss die Identifikation von Arbeitssituationen und Personenkreisen sowie Gestaltungskriterien sein, die ein Unternehmen dazu befähigt, geeignete Konfigurationen von RPM zu gestalten (vgl. K. Murphy und A. DeNisi 2023, S. 143; K. R. Murphy 2020, S. 13).

Zunächst wird der grundlegende Wirkungsmechanismus beschrieben. Aufgrund der Natur des RPM entsteht eine Situation, die einen SV unter den Beteiligten ermöglicht und fördert (vgl. M. Garcia und A. Tor 2007, S. 95; J. Luft 2016, S. 75–80). Wettbewerb unter den Beteiligten und der Wunsch zu gewinnen ist eine anerkannte Folge dieser Vergleichsprozesse. Die Wettbewerbsposition wird dabei konsequent entlang einer definierten Performancedimension geordnet. Die Aktivierung kann als Ergebnis aus der Verarbeitung relativer Performance-Informationen (RPI) verstanden werden (vgl. M. Garcia et al. 2013, S. 634; B. Holmstrom 1982, S. 325; P. Lazear und S. Rosen 1981, S. 641; J. Luft 2016, S. 78–80). Wettbewerb wird durch eine Konstellation charakterisiert, in der die Zielerreichung der Teilnehmenden negativ korreliert. Jede Position in der Verteilung, jeder Rang kann nur einmal besetzt werden (vgl. E. Dechenaux et al. 2015, S. 613; N. Eichhorn 2016, S. 105). In Anlehnung an das Rangordnungsparadigma lässt sich die Verhaltens- und Motivationswirkung von Wettbewerb in drei zentrale Bestandteile gliedern. Erstens existieren ein oder mehrere Wettbewerber, die in einem SV zueinander stehen. Zweitens gibt es ein klar definiertes Wettbewerbsziel, das den Gegenstand der Konkurrenz bildet. Drittens führt das Zusammenspiel dieser Elemente zu wettbewerbsorientierten Verhaltensweisen (vgl. M. Garcia et al. 2013, S. 634f.; M. Garcia und A. Tor 2007, S. 95, 104–106; A. Tran und R. Zeckhauser 2012, S. 645). Ökonomische Ansätze erkennen das Potenzial von Wettbewerb im Kontext des MC bereits seit Langem an, gehen aber davon aus, dass Verhaltenswirkungen nur bei der Kopplung eines RPM-Systems an monetäre Anreize auftreten (vgl. E. Bonner und B. Sprinkle 2002, S. 303–308; B. Holmstrom 1982, S. 324; J. Luft 2016, S. 75; C. Schnieder 2022, S. 74; D. Tafkov 2013, S. 339; A. Tran und R. Zeckhauser 2012, S. 645). Seit den 2000er-Jahren wird diese Annahme zunehmend durch sozial-psychologische Perspektiven hinterfragt. Es wird postuliert, dass Wettbewerb auch um nicht-monetäre Belohnungen wie Überlegenheit in Fähigkeiten, Stolz, Anerkennung oder allgemein positive Emotionen wirksam sein kann (vgl. N. Eichhorn 2016, S. 102; S. Kramer et al. 2016, S. 16; J. Luft 2016, S. 77–80; C. Schnieder 2022, S. 73–75; D. Tafkov 2013, S. 339). Wettbewerbsfähigkeit lässt sich in zwei Kategorien unterteilen, eine aufgabenorientierte und eine rivalitätsorientierte. Die aufgabenorientierte Wettbewerbsfähigkeit richtet den Fokus auf die tatsächliche Performance und fördert die Verbesserung von Fähigkeiten oder die Steigerung der Anstrengung. Im Gegensatz dazu zielt die rivalitätsorientierte Wettbewerbsfähigkeit auf eine bessere Position in der Performanceverteilung (z. B. einen höheren Rang) ab, was auch durch dysfunktionales Verhalten wie Sabotage erreicht werden kann (vgl. N. Eichhorn 2016, S. 105; M. Garcia et al. 2013, S. 634; M. Garcia und A. Tor 2007, S. 97; F. Hartmann und P. Schreck 2018, S. 366). Insgesamt ist die Wirkung von Wettbewerb auf das Verhalten, die Motivation und die Performance weder eindeutig noch immer positiv (vgl. E. Bonner und B. Sprinkle 2002, S. 309, 329; N. Kluger und A. DeNisi 1996, S. 258f.; J. Luft 2016, S. 78; H. Song et al. 2018, S. 2629; D. Tafkov 2013, S. 330). Die zentrale Herausforderung besteht darin, die gegenläufigen Effekte von RPM-Systemen zu verstehen und bei der Gestaltung von Anreizsystemen positive Wirkungen zu maximieren, während negative minimiert oder gänzlich vermieden werden (vgl. S. Blader et al. 2019, S. 721; G. Charness et al. 2014, S. 38; F. Hartmann und P. Schreck 2018, S. 364; J. Luft 2016, S. 80).

Diese Arbeit beleuchtet zwei häufig verwendete theoretische Ansätze, die das Potenzial des Wettbewerbsmechanismus im unternehmerischen Kontext erklären. Einerseits wird der ökonomische Ansatz der AT betrachtet (vgl. R. J. Akerlof und R. T. Holden 2012, S. 289f.; T. Ball et al. 2020, S. 1507; E. Dechenaux et al. 2015, S. 610; A. Dye 1992, S. 27f.; R. Frederickson 1992, S. 649–651; S. Goebel und E. Weißenberger 2017, S. 191f.; L. Hannan et al. 2008, S. 898; B. Holmstrom 1982, S. 334–338; P. Lazear und S. Rosen 1981, S. 841; J. Luft 2016, S. 75–79), andererseits die TSV (vgl. G. Birnberg et al. 2006, S. 121–123; G. Charness et al. 2014, S. 38f.; Y. Chen et al. 2010, S. 1359; N. Eichhorn 2016, S. 100–103; L. Hannan et al. 2008, S. 887–893; S. Kramer et al. 2016, S. 16–18; B. Lount und W. Phillips 2007, S. 214; J. Luft 2016, S. 75–78; S. Murthy und A. Schafer 2011, S. 160–163; D. Tafkov 2013, S. 329–333).

W. Thorngate 1976 postuliert in seiner Theorie der angemessenen Komplexität, dass eine Theorie zum Verhalten im sozialen Kontext kein Gleichgewicht zwischen den Kriterien Allgemeinheit, Genauigkeit und Einfachheit herstellen kann. Allgemeinheit bedeutet universelle Anwendbarkeit, Genauigkeit, die Eignung für die konkrete Anwendung, und Einfachheit, die leichte Verständlichkeit (vgl. W. Thorngate 1976, S. 406; E. Weick 2015, S. 54f.). E. Weick (2015, S. 54–64) entwickelt aus diesen Gedanken die Komplexitätsuhr (Abb. 2), die zeigt, dass beim Streben nach zwei Tugenden eine dritte aufgegeben werden muss.

Die AT entspricht der Sechs-Uhr-Forschung und kombiniert die Kriterien Einfachheit und Genauigkeit (vgl. E. Weick 2015, S. 55). Im Mittelpunkt der AT steht die hierarchische Beziehung zwischen einem Prinzipal und einem Agenten, die durch Zielkonflikte und asymmetrische Informationen geprägt ist (vgl. P. Baker 1992, S. 599f.; T. Ball et al. 2020, S. 1507; S. Goebel und E. Weißenberger 2017, S. 191; J. Luft 2016, S. 75). In der Beschreibung dieses engen Anwendungsbereiches kann die Theorie als genau bezeichnet werden (vgl. S. Goebel und E. Weißenberger 2017, S. 192).

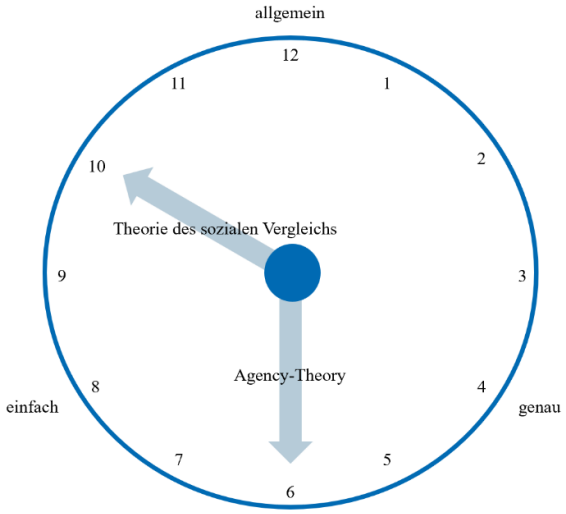

Abbildung 2: Einordnung der Agency-Theory und der Theorie des sozialen Vergleichs in die Komplexitätsuhr.

(Quelle: leicht verändert entnommen aus E. Weick 2015, S. 55)

Der Prinzipal beauftragt den Agenten mit einer Aufgabe, hat jedoch unzureichende Informationen über dessen Fähigkeiten und Verhalten (vgl. R. J. Akerlof und R. T. Holden 2012, S. 392; P. Baker 1992, S. 599; A. Dye 1992, S. 27). Es wird angenommen, dass der Agent rational und risikoscheu handelt, motiviert von einer opportunistischen Maximierung seiner Nutzenfunktion (vgl. E. Bonner und B. Sprinkle 2002, S. 308; S. Goebel und E. Weißenberger 2017, S. 192). Im Rahmen des OoC-Framework kann Moral Hazard (MH) als mangelnde Motivation interpretiert werden. Der Agent kann nach Vertragsabschluss opportunistisch handeln, ohne dass der Prinzipal über ausreichend informative Beobachtungsgrößen verfügt (vgl. P. Baker 1992, S. 600; B. Holmstrom 1982, S. 324, 338; P. Lazear und S. Rosen 1981, S. 841). Die AT empfiehlt Anreizvergütungsvereinbarungen einschließlich RPM, um sicherzustellen, dass Agenten im Einklang mit den Prinzipalzielen handeln (vgl. R. J. Akerlof und R. T. Holden 2012, S. 289; T. Ball et al. 2020, S. 1507; A. Dye 1992, S. 28; R. Frederickson 1992, S. 647; S. Goebel und E. Weißenberger 2017, S. 191). Der Ansatz der AT vernachlässigt einen Großteil der Komplexität sozial geprägter Unternehmenssysteme, weshalb es notwendig ist, ergänzend einen verhaltenstheoretischen Blickwinkel einzubeziehen, um das Verständnis der Auswirkungen von RPM ins Allgemeine zu erweitern (vgl. G. Birnberg et al. 2006, S. 113f.; S. Goebel und E. Weißenberger 2017, S. 192).

Die TSV stellt die Zehn-Uhr-Forschung dar und vereint die Kriterien Einfachheit und Allgemeinheit (vgl. E. Weick 2015, S. 55). Abb. 3 veranschaulicht den SV-Prozess, der als unvermeidbares Element sozialer Interaktion gilt und auch im Unternehmenskontext allgegenwärtig ist (vgl. J. Brown et al. 2007, S. 59f.; N. Eichhorn 2016, S. 102; M. Garcia et al. 2013, S. 634).

L. Festinger (1954) (S. 117–120) beschreibt das menschliche Grundbedürfnis, die eigenen Fähigkeiten und die daraus resultierende Performance aus einem Grundzustand des Selbst heraus bewerten zu wollen. Fehlt ein objektiver Maßstab, greift das Individuum auf SV zur Bedürfnisbefriedigung zurück (vgl. J. Brown et al. 2007, S. 59; P. Buunk et al. 2005, S. 4; N. Eichhorn 2016, S. 102; L. Festinger 1954, S. 117–120; L. Hannan et al. 2008, S. 896; S. Murthy und A. Schafer 2011, S. 162; C. Schnieder 2022, S. 75). RPM stellt in diesem Zusammenhang Informationen zur Verfügung, die eine solche Selbstbewertung in Relation zu einer Vergleichsgruppe ermöglichen (vgl. N. Eichhorn 2016, S. 102; C. Schnieder 2022, S. 75; J. Suls et al. 2002, S. 159). Die Ergebnisse des Vergleichsprozesses beeinflussen durch das *„Nachdenken über Informationen bezüglich einer oder mehrere andere Personen im Verhältnis zum Selbst“* (J. Brown et al. 2007, S. 60) das Selbstbild und somit auch den Selbstwert. Infolgedessen können sowohl positive als auch negative Emotionen und affektive Reaktionen hervorgerufen werden (vgl. J. Brown et al. 2007, S. 60; E. Carroll und D. Marginson 2021, S. 1; S. Murthy und A. Schafer 2011, S. 163; C. Schnieder 2022, S. 75; J. Suls et al. 2002, S. 159). Das Streben nach einem stabilen und positiven Selbstbild kann SV als Quelle für Wettbewerbsverhalten identifizieren. Im Allgemeinen werden Anstrengungs- und Motivationseffekte durch den Wettbewerb angeregt (vgl. J. Brown et al. 2007, S. 60; E. Carroll und D. Marginson 2021, S. 1; N. Eichhorn 2016, S. 102; M. Garcia et al. 2013, S. 634f.; M. Garcia und A. Tor 2007, S. 95; L. Hannan et al. 2008, S. 896; D. Mahlendorf et al. 2014, S. 333; S. Murthy und A. Schafer 2011, S. 162; H. Smith 2000, S. 174f.; D. Tafkov 2013, S. 329f.). Die Entscheidung für einen Aufwärts- oder Abwärtsvergleich ist ein subjektiver kognitiver Prozess. Beide Vergleichsrichtungen können sowohl positive als auch negative affektive Reaktionen hervorrufen, die unterschiedliche Verhaltenseffekte auslösen (vgl. J. Brown et al. 2007, S. 60f.; P. Buunk et al. 2005, S. 8–10; Y. Chen et al. 2010, S. 1359; S. Murthy und A. Schafer 2011, S. 163; H. Smith 2000, S. 175; J. Suls et al. 2002, S. 161; D. Tafkov 2013, S. 329). In diesem Kontext stehen sich assimilative und kontrastive Selbstbewertung gegenüber (vgl. J. Brown et al. 2007, S. 59f.; N. Eichhorn 2016, S. 103; L. Festinger 1954, S. 120–130; H. Smith 2000, S. 178). Die TSV geht ursprünglich von einem unidirektionalen Drang nach oben aus, einem assimilativen Effekt des Aufwärtsvergleichs. Das Ziel ist die Annäherung an überlegene Individuen und der Selbstverbesserungsimpuls resultiert aus Bewunderung und Inspiration (vgl. G. Birnberg et al. 2006, S. 121; J. Brown et al. 2007, S. 60; E. Carroll und D. Marginson 2021, S. 1; N. Eichhorn 2016, S. 102; L. Festinger 1954, S. 124f.; M. Garcia et al. 2013, S. 634; M. Garcia und A. Tor 2007, S. 95; B. Lount und W. Phillips 2007, S. 215; S. Murthy und A. Schafer 2011, S. 162; H. Smith 2000, S. 180–186; J. Suls et al. 2002, S. 161). Im Gegensatz dazu kann eine kontrastive Einstellung beim Aufwärtsvergleich Scham und Neid hervorrufen, was entweder zu einem Abbruch des Vergleichs oder zu dysfunktionalen, sabotierenden Verhaltensweisen führt (vgl. E. Carroll und D. Marginson 2021,

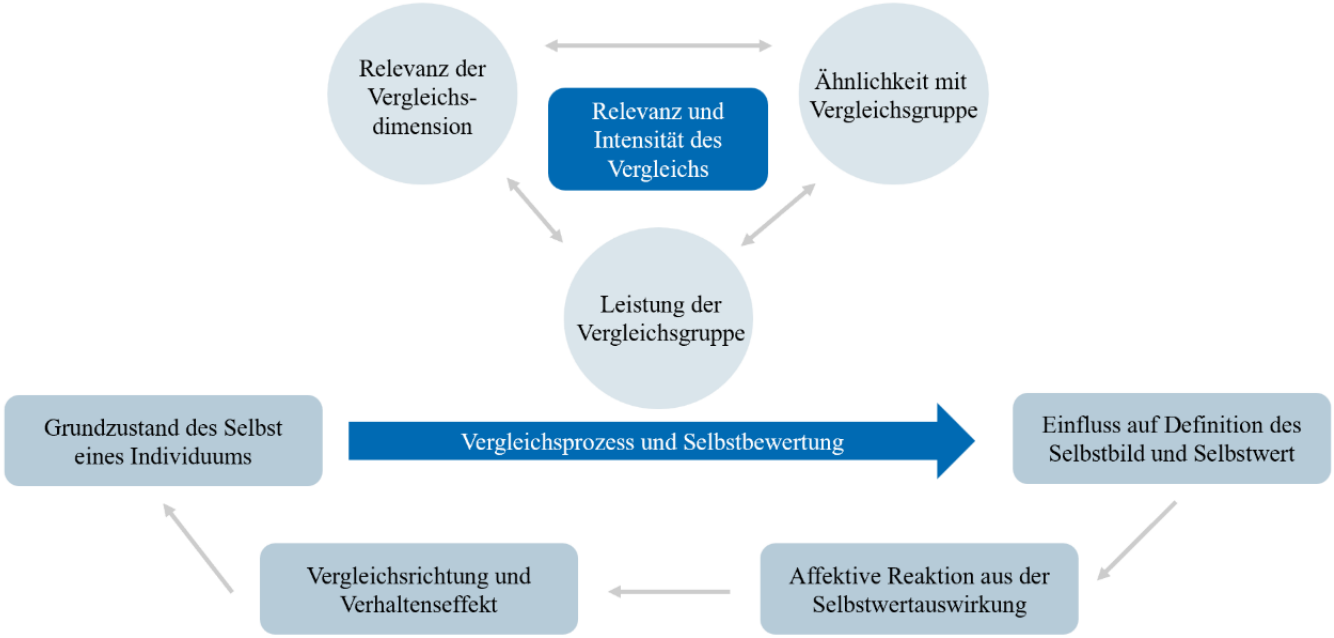

Abbildung 3: Prozess des sozialen Vergleichs.

S. 12; G. Charness et al. 2014, S. 39; N. Eichhorn 2016, S. 103; L. Festinger 1954, S. 129; M. Garcia et al. 2013, S. 635; H. Smith 2000, S. 180–186; J. Suls et al. 2002, S. 161). Der Abwärtsvergleich kann als Selbstwertschutz durch Kontrastbildung genutzt werden, indem der Vergleich mit leistungsschwächeren Personen das eigene Selbstbild stärkt (vgl. G. Birnberg et al. 2006, S. 121; J. Brown et al. 2007, S. 60; E. Carroll und D. Marginson 2021, S. 12; N. Eichhorn 2016, S. 103; A. Mohnen et al. 2008, S. 633; H. Smith 2000, S. 187–192). Assimilativ birgt er jedoch für Leistungsstarke die Gefahr, Ängste und Sorgen über einen möglichen Leistungsabfall zu wecken (vgl. E. Carroll und D. Marginson 2021, S. 12; G. Charness et al. 2014, S. 39; N. Eichhorn 2016, S. 103; M. Garcia et al. 2013, S. 635; H. Smith 2000, S. 187–192). Zusammenfassend lässt sich festhalten, dass die gewählte Vergleichsrichtung die daraus resultierenden Verhaltens-, Anstrengungs- und Performanceeffekte nicht eindeutig sind. Sie hängen sowohl von den individuellen Prädispositionen als auch von der Umgebung ab, in der die Selbstbewertung stattfindet (vgl. J. Brown et al. 2007, S. 60–69; N. Eichhorn 2016, S. 103; M. Garcia et al. 2013, S. 634–637; L. Hannan et al. 2008, S. 896; H. Smith 2000, S. 180–192).

Das Self-Evaluations-Maintenance-Modell von A. Tesser 1988 ist eine wichtige Erweiterung der TSV. Es führt drei Kriterien ein, die bestimmen, wann Individuen am SV teilnehmen und wann das resultierende wettbewerbsorientierte Verhalten besonders ausgeprägt ist (vgl. H. Beach und A. Tesser 1995, S. 148; P. Buunk et al. 2005, S. 10; M. Garcia und A. Tor 2007, S. 182–206; D. Tafkov 2013, S. 330; A. Tesser 1988, S. 148f.; A. Tesser et al. 1989, S. 442f.). Erstens ist die psychologische Nähe zur Vergleichsperson entscheidend (vgl. M. Garcia et al. 2013, S. 637; G. Harkins und M. Jackson 1985, S. 457; D. Tafkov 2013, S. 330; A. Tesser 1988, S. 182). In Anlehnung an Festingers Ähnlichkeitshypothese (vgl. L. Festinger 1954, S. 121f.) zeigt sich, dass der SV-Prozess umso relevanter wird, je ähnlicher die Vergleichsgruppe hinsichtlich fähigkeitsdeterminierender Attribute wie der Persönlichkeit, individuelle Ziele und sozialen Erfahrungen ist. Im Unternehmenskontext bedeutet dies insb. eine ähnliche Berufsidentifikation und Aufgabensituation (vgl. N. Eichhorn 2016, S. 103; G. Harkins und M. Jackson 1985, S. 457; D. Mahlendorf et al. 2014, S. 334; D. Tafkov 2013, S. 330; A. Tesser 1988, S. 182). Das Individuum kümmert sich um sein soziales Image, wie es von anderen in der Vergleichsgruppe wahrgenommen wird und versucht, Andere positiv zu beeindrucken (vgl. H. Beach und A. Tesser 1995, S. 145; G. Birnberg et al. 2006, S. 123; G. Harkins und K. Szymanski 1989, S. 935; A. Hogg und J. Terry 2000, S. 121–124; R. Leary und M. Kowalski 1990, S. 34f.; C. Schnieder 2022, S. 75). Zweitens ist die Relevanz der Vergleichsdimension wichtig. Die gemessene Performance im RPM-System muss für das Selbstkonzept des Einzelnen von signifikanter Bedeutung sein, andernfalls bleibt der SV-Prozess aus (vgl. N. Eichhorn 2016, S. 103; M. Garcia et al. 2013, S. 637; M. Garcia und A. Tor 2007, S. 95; D. Tafkov 2013, S. 330; A. Tesser 1988, S. 183). Drittens spielt die Positionierung innerhalb der Vergleichsgruppe eine zentrale Rolle. Die Spannbreite der Leistungsverteilung und die relative Platzierung des

Individuums beeinflussen die Wahl der Vergleichsrichtung sowie die affektiven Verhaltenseffekte (vgl. M. Garcia et al. 2013, S. 637; D. Mahlendorf et al. 2014, S. 333; A. Tesser 1988, S. 182; A. Tesser et al. 1989, S. 442).

3 Potenzial und Grenzen des Turnieransatzes in der Unternehmenssteuerung

Unter Anwendung der AT hat sich in der wirtschaftswissenschaftlichen Literatur ein PM-Wettbewerbsmodell als MCS etabliert, das Ranglisten-Turniersystem (vgl. E. Dechenaux et al. 2015, S. 610). Analytische Modelle orientieren sich bis heute an der bewährten Grundlage solcher Turnierformalismen im Managementbereich, die maßgeblich von P. Lazear und S. Rosen 1981 sowie B. Holmstrom 1982 geprägt wurden (vgl. T. Ball et al. 2020, S. 1507f.; E. Dechenaux et al. 2015, S. 619f.; R. Frederickson 1992, S. 649f.; B. Holmstrom 1982, S. 324; P. Lazear und S. Rosen 1981, S. 841f.; J. Luft 2016, S. 75). Das Turniermodell wird typischerweise als ordinales RPM-System interpretiert, bei dem Teilnehmende im Wettbewerb um einen Preis antreten. Dieser Preis ist meist eine explizite monetäre Belohnung, wie ein Bonus oder eine variable Vergütung. In der klassischen Form wird ein einzelner Preis an die ranghöchste und damit leistungsstärkste Person vergeben (vgl. R. J. Akerlof und R. T. Holden 2012, S. 290; E. Dechenaux et al. 2015, S. 610–614; A. Dye 1992, S. 27f.; B. Holmstrom 1982, S. 325; P. Lazear und S. Rosen 1981, S. 844). Turniere ermöglichen eine effizientere Risikoverteilung bei risikoaversen Agenten und gelten als kosteneffiziente Alternative. Der Mechanismus der Turniere setzt automatisch Standards innerhalb der Turniergruppe und passt diese dynamisch an veränderte Rahmenbedingungen an. Dadurch zeichnet sich RPM durch eine hohe Flexibilität aus (vgl. E. Dechenaux et al. 2015, S. 619f.; B. Holmstrom 1982, S. 338; P. Lazear und S. Rosen 1981, S. 841; J. Nalebuff und E. Stiglitz 1983, S. 21).

Um das ökonomische Argument der grundlegenden Nützlichkeit von Turniersystemen im Rahmen von MC näher zu erläutern, werden zunächst allgemeine Modellkomponenten eingeführt. Ein Prinzipal beauftragt einen Agenten mit der Erfüllung einer Aufgabe, deren Ergebnis in Form der beobachtbaren individuellen Performance x_i gemessen wird. Die zugrunde liegende Performancefunktion lautet:

$$x_i = e_i + \mathscr{E}_c + \mathscr{E}_i \tag{1}$$

Dabei setzt sich x_i aus drei Komponenten zusammen. Erstens dem vom Agenten privat gewählten Anstrengungsniveau e_i, das vollständig unter seiner Kontrolle steht. Zweitens der allgemeinen Unsicherheitskomponente ε_c, die alle am Turnier beteiligten Agenten gleichermaßen betrifft und Einflüsse wie konjunkturelle Schwankungen, regulatorische Änderungen oder marktspezifische Preisvolatilitäten umfasst. Drittens der individuellen, idiosynkratischen Unsicherheitskomponente ε_i, die durch spezifische Eigenschaften des Agenten geprägt ist. Darunter fallen seine Fähigkeiten, die Kosten für Anstrengung und Qualifikation sowie die Eignung zur Erfüllung der Aufgabe (vgl. R. J. Akerlof und R. T. Holden 2012, S. 292; E. Dechenaux et al. 2015, S. 611; R. Frederickson 1992, S. 649f.; P. Lazear und S. Rosen 1981, S. 843; J. Nalebuff und E. Stiglitz 1983, S. 21). Dem Agenten entstehen bei der Erbringung von Anstrengungseinheiten nicht-monetäre Kosten, die durch die Kostenfunktion der Formel (2) beschrieben werden.

$$c(e_i) = \frac{1}{2} * k * e_i^2 \tag{2}$$

Diese Funktion berücksichtigt positive und steigende Grenzkosten, um den Einsatz von Anstrengung zu begrenzen. Ab einem bestimmten Punkt wird es für den rationalen Agenten aufgrund seiner Kosten-Nutzen-Abwägung ineffizient und unattraktiv, weiter in Anstrengung zu investieren. Der Parameter k bestimmt die individuelle Kostenempfindlichkeit und erlaubt heterogene Kostenfunktionen der Agenten. Höhere Kosten für ein spezifisches Anstrengungsniveau deuten auf eine geringere Eignung eines Agenten im Vergleich zu einem mit niedrigeren Kosten hin (vgl. R. J. Akerlof und R. T. Holden 2012, S. 291f.; E. Dechenaux et al. 2015, S. 613; B. Holmstrom 1982, S. 325; P. Lazear und S. Rosen 1981, S. 843f.) Das zentrale MH-Problem für den Prinzipal besteht darin, dass er das Anstrengungsniveau e_i des Agenten nicht direkt beobachten kann, jedoch ein strategiekonformes Niveau anstrebt (vgl. T. Ball et al. 2020, S. 1510; A. Dye 1992, S. 27f.; R. Frederickson 1992, S. 649; B. Holmstrom 1982, S. 338). Die Gesamtperformance der Turnierteilnehmer ergibt sich aus der Summe der individuellen Performance aller beteiligten Agenten, wobei in der Literatur häufig vereinfacht von zwei Agenten i und j, ausgegangen wird. Die Gesamtperformance X lässt sich somit wie folgt darstellen (vgl. R. J. Akerlof und R. T. Holden 2012, S. 292; E. Dechenaux et al. 2015, S. 611; R. Frederickson 1992, S. 649f.; P. Lazear und S. Rosen 1981, S. 843):

$$X = x_i + x_j \tag{3}$$

Bietet der Prinzipal nun einen Turniervertrag in der o. g. Grundform an, ergibt sich seine Auszahlung als Gesamtperformance der Agenten abzüglich der jeweiligen fixen Grundentlohnung L und der Finanzierung des variablen Turnierpreises W. Den Preis erhält nur der Agent mit der höchsten Einzelperformance. Für die Vertragsgestaltung nutzt der Prinzipal die Gesamtperformance X, die eine Rangfolge der Einzelperformances x_i und x_j ermöglicht (vgl. E. Dechenaux et al. 2015, S. 619). Betrachtet man in diesem Zusammenhang den eigennützig maximierenden Agenten i, lässt sich die Nutzenfunktion wie in Formel (4) darstellen.

$$U_i = \mathbb{P}\left(x_i > x_j\right) * W + L - c(e_i) \tag{4}$$

Die Gewinnwahrscheinlichkeit $\mathbb{P}\left(x_i > x_j\right)$ hängt davon ab, dass die Performance von Agent i die von Agent j über-

steigt. Im Rahmen des Modells wird diese Wahrscheinlichkeit durch die relativen Investitionen von Agent i in das Anstrengungsniveau e_i determiniert, mit dem Ziel, den Preis W zu gewinnen. Das Streben aller Agenten nach dem Turniergewinn impliziert einen positiven Anstrengungseffekt (vgl. E. Dechenaux et al. 2015, S. 612–614). Soweit zum allgemeinen Wettbewerbsmodell.

Wie in Formel (1) dargestellt, ist das Signal x_i durch zwei Unsicherheitskomponenten in seiner Aussagekraft über die tatsächliche Anstrengung e_i eingeschränkt. Mittels der Relativierung im Turnier und der Verwendung des Signal X gelingt es, die gemeinsame Unsicherheitskomponente ε_c aus der Performancefunktion herauszufiltern, sodass lediglich ε_i übrig bleibt. Im Turniervertrag liefert die Performance eines Agenten daher wertvolle Informationen über alle anderen und erlaubt dem Prinzipal einen besseren Rückschluss auf die Anstrengung des einzelnen Agenten (vgl. R. Frederickson 1992, S. 647–651; B. Holmstrom 1982, S. 324; J. Nalebuff und E. Stiglitz 1983, S. 21–24). Das Signal-Rausch-Verhältnis, also der Informationsgehalt und damit die Wirkung und Nützlichkeit von RPM im Turnier, steigt mit Zunahme des Anteils der gemeinsamen Unsicherheit an der Gesamtunsicherheit (vgl. R. Frederickson 1992, S. 650f.). Der Einsatz von RPM reduziert MH-Auswirkungen und verbessert die Risikoallokation, indem er durch erhöhte gemeinsame Unsicherheit mehr über die Anstrengung des Agenten offenbart, was die Motivation steigert. Dies fördert sowohl die Auszahlung des Prinzipals als auch die Unternehmensperformance und ermöglicht eine effizientere Steuerung im Rahmen der MC. Voraussetzung für diese Argumentation ist jedoch eine ausreichende gemeinsame Unsicherheitskomponente, die ein signifikantes Niveau erreicht, um einen Anstrengungseffekt zu erzielen (vgl. R. J. Akerlof und R. T. Holden 2012, S. 290; R. Frederickson 1992, S. 649–651; B. Holmstrom 1982, S. 235–238).

Grenzen der grundsätzlichen Nützlichkeit von Turniersystemen werden bereits durch die häufig strengen Annahmen der Modelle gesetzt. Der Untersuchungsrahmen basiert in der Regel auf Agenten gemäß AT (Kapitel 2.2), beschränkt sich auf eine einzelne Periode und geht von homogenen Fähigkeiten der Teilnehmer aus (vgl. R. J. Akerlof und R. T. Holden 2012, S. 290; E. Dechenaux et al. 2015, S. 613; B. Holmstrom 1982, S. 334; N. Janakiraman et al. 1992, S. 55). Zudem kann die Wettbewerbsstruktur erheblich von der Grundform abweichen. Wichtige Variablen in diesem Kontext sind die Anzahl der Teilnehmer, die Anzahl der Preise, die Einführung heterogener Agenten sowie die Berücksichtigung von Sanktionen für den Turnierletzten (vgl. E. Dechenaux et al. 2015, S. 621; A. Dye 1992, S. 28; J. Luft 2016, S. 79). Studien zeigen eine negative Beziehung zwischen der Anstrengung der Agenten und der Gruppengröße im Turnier, insb. in Abhängigkeit von idiosynkratischen Risikokomponenten und deren Verteilung. Die Auswirkungen individueller Verhaltensweisen bei steigender Spieleranzahl sind uneinheitlich und führen zu einer für das Unternehmen nachteiligen Volatilität (vgl. R. J. Akerlof und R. T. Holden 2012, S. 302–305; E. Dechenaux et al. 2015, S. 622). Eine breite Performancespanne aufgrund unterschiedlicher Fähigkeiten macht eine Investition in Anstrengung bei leistungsschwachen Teilnehmern zunehmend unrentabel und ruft Entmutigungseffekte hervor, sodass diese weniger Anstrengung investieren oder aufgeben (vgl. P. Casas-Arce und A. Martínez-Jerez 2009, S. 1306; E. Dechenaux et al. 2015, S. 622f.; J. Luft 2016, S. 79f.). Im Turnierrahmen kann diesem Phänomen durch Anpassung der Preisanzahl und -werte entgegengewirkt werden. Schwellenwertpreise mit absteigendem Wert können neue Motivationseffekte erzeugen. Die optimale Preisstruktur hängt von mehreren Faktoren ab, darunter die Fähigkeitsheterogenität, die Anstrengungskostenfunktionen und der Grad der Risikoaversion der Agenten (vgl. R. J. Akerlof und R. T. Holden 2012, S. 302–305; E. Dechenaux et al. 2015, S. 624). Die Erweiterung der betrachteten Aufgabe von Eindimensionalität auf Mehrdimensionalität verdeutlicht zusätzliche Schwierigkeiten in den Modellen. Ein Agent kann seine Anstrengung so aufteilen, dass er sich auf Aufgabendimensionen konzentriert, bei denen seine Fähigkeiten am stärksten ausgeprägt sind, anstatt eine aus Unternehmenssicht optimale Verteilung zu verfolgen. Darüber hinaus kann der starke Wettbewerb im Turnier durch Ablenkungseffekte wie Verwirrung und Stress den Fokus des Agenten von der Aufgabe auf den Wettbewerb selbst verschieben (vgl. A. Dye 1992, S. 27f.; J. Luft 2016, S. 78–80). Die meisten analytischen Turniermodelle der Literatur basieren auf einer Prinzipal-Agenten-Beziehung wie in Kapitel 2.2 beschrieben und fokussieren sich auf hierarchische Verhältnisse zur Steuerung von Managern und Führungskräften. Dabei erfolgt die Relativierung innerhalb eines ähnlichen Wettbewerbsumfelds, in dem die Vergleichsobjekte aus einer gleichen oder zumindest hinreichend ähnlichen Branche stammen (vgl. T. Ball et al. 2020, S. 1507; A. Dye 1992, S. 45; P. Lazear und S. Rosen 1981, S. 841; J. Luft 2016, S. 75).

Eine wesentliche Schwierigkeit von Turniersystemen liegt in der Übertragung ihrer Steuerungswirkung auf horizontale Beziehungen und den unternehmensinternen Kontext. Vor diesem Hintergrund soll das Konzept des Forced-Ranking (FR) als ebensolcher Versuch untersucht werden (vgl. J. Luft 2016, S. 75). Das FR-Modell wurde 1981 von Jack Welch, dem damaligen Chief-Executive-Officer von General Electric, eingeführt (vgl. H. Beth und M. Sasse 2008, S. 35; A. Wijayanti et al. 2024, S. 9f.). Welch propagierte, dass sein Einsatz dieses RPM-Systems in nur fünf Jahren zu einer Umsatzsteigerung von 70 Milliarden auf 130 Milliarden USD führte (J. Alsever 2008, Abs. 1–4). FR hat sich seitdem in der Praxis zu einer weitverbreiteten Methode entwickelt. Es wird angenommen, dass ein erheblicher Teil der Unternehmen, darunter ein Drittel der Fortune 500, ein System anwendet, das dem FR-Modell nachempfunden ist (vgl. J. Alsever 2008, Abs. 1–4; H. Beth und M. Sasse 2008, S. 35; A. Long 2021, Abs. 1–2). Eine Befragung des *Jacksonville Business Journal* ergab, dass 60 % der Teilnehmer eine Art FR nutzen, darunter prominente Unternehmen wie General Electric, Yahoo!, Microsoft, Intel, Ford, Goldman Sachs und American Express (vgl. S. Stewart et al. 2010, S. 169). Aufgrund seiner praktischen Relevanz hat es inzwischen auch als Untersuchungsobjekt Eingang in die RPM-Literatur gefunden.

FR folgt dem Turnieransatz und erstellt eine Rangliste aller beteiligten Mitarbeitenden, wobei es beispielsweise als 20-70-10-Turnier interpretiert werden kann. Dabei werden die Mitarbeitenden in drei Gruppen eingeteilt: die obersten 20 % der Leistungsstarken, die 70 % im Mittelfeld und die unteren 10 % der Leistungsschwachen. Die leistungsstarken Mitarbeiter erhalten Belohnungen wie Boni, Gehaltserhöhungen oder Beförderungen, während die Mitarbeitenden im Mittelfeld neutral behandelt werden. Die leistungsschwachen Mitarbeitenden sind einer Sanktion ausgesetzt, die von finanziellen Einbußen über Schulungsprogramme bis hin zur Entlassung reichen kann, wie es ursprünglich unter dem Konzept *„Rank and Yank“* (A. Wijayanti et al., 2024, S. 1) gehandhabt wurde (vgl. J. Alsever 2008, Abs. 1–4; H. Beth und M. Sasse 2008, S. 35; E. Cardinaels und C. Feichter 2023, S. 15; A. Long 2021, Abs. 1–2; S. Stewart et al. 2010, S. 168; A. Wijayanti et al. 2024, S. 3). Das FR-Modell basiert auf der Annahme, dass die Performance der Mitarbeitenden normalverteilt ist (vgl. S. Stewart et al. 2010, S. 170), was durch die blaue Glockenkurve in Abb. 4 veranschaulicht wird.

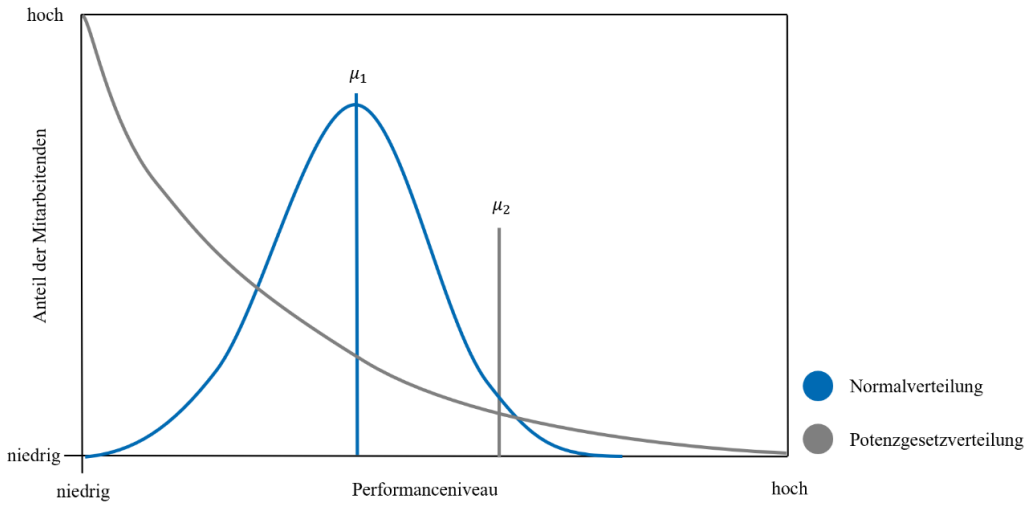

Abbildung 4: Vergleich der Normalverteilung mit der Potenzgesetzverteilung für die Performance unter den Mitarbeitenden.

(Quelle: leicht verändert entnommen aus H. Aguinis und K. J. Bradley 2015, S. 162.)

Das Ziel von FR besteht darin, leistungsstarke Mitarbeitende zu bestätigen und zu fördern sowie die mittlere Performance zu stabilisieren oder in Richtung der Belohnungszone zu lenken. Die leistungsschwächsten Mitarbeitenden sollen entweder durch gezielte Förderung aus der Talsohle befördert oder durch Kündigung aus dem Unternehmen entfernt werden. Das Unternehmen verstärkt den Wettbewerb, nutzt eine Kosten-Nutzen-Überlegung und verschiebt die *„Vitalitätskurve“* (D. Gill et al. 2019, S. 494) nach rechts, wodurch die Gesamtperformance des Unternehmens theoretisch verbessert wird (vgl. E. Cardinaels und C. Feichter 2023, S. 15; D. Gill et al. 2019, S. 494; A. Wijayanti et al. 2024, S. 2). Neben den Performanceeffekten soll FR durch die vollständige Differenzierung entlang der Rangliste eine nachsichtige Zielsetzung und Bewertung sowie die Zentralisation mehrere Leistungsträger auf einem Niveau vermeiden (vgl. S. Stewart et al. 2010, S. 171; A. Wijayanti et al. 2024, S. 2). Insgesamt werden FR die Vorteile einer verbesserten Bewertungsgenauigkeit und eines wettbewerbsbasierten Motivationseffekts zugeschrieben. Aus der Forschungsperspektive zeigt FR eine Reihe von Nachteilen, die das ursprüngliche Ziel des Turniermodells verfehlen und die einfache Mechanik infrage stellen. Statt Performancesteigerungen zu erzielen, führen FR-Systeme häufig zu verzerrten Informationen über die tatsächliche Anstrengung und Performancefähigkeit der Mitarbeitenden. FR ist in diesem Kontext durch Emotions- und Stresseffekte geprägt (vgl. E. Cardinaels und C. Feichter 2023, S. 18f.; S. Stewart et al. 2010, S. 172–176; A. Wijayanti et al. 2024, S. 2–4). Negative affektive Reaktionen wirken sich insb. bei langfristiger Anwendung nachteilig auf die Gesamtperformance aus (vgl. T. Baldwin et al. 2013, S. 360; E. Cardinaels und C. Feichter 2023, S. 19). Neben der Angst vor Bestrafung oder dem Entzug von Belohnungen entstehen Gefühle der Ungerechtigkeit und Eifersucht, die zu dysfunktionalem Verhalten wie Leistungsverweigerung oder Sabotage führen (vgl. J. Alsever 2008, Abs. 1–4; T. Baldwin et al. 2013, S. 371f.; E. Cardinaels und C. Feichter 2023, S. 15f.; S. Stewart et al. 2010, S. 176; A. Wijayanti et al. 2024, S. 3f.).

Die Kritik an der Turnieranwendung lässt sich nicht nur auf das Scheitern von Modellannahmen, sondern auch auf eine fehlerhafte Annahme der Performanceverteilung zurückführen. Während FR und viele PM- bzw. RPM-Methoden in der Management- und Organisationsforschung auf der Normalverteilung basieren, zeigen Untersuchungen von zeigen Untersuchungen von H. Aguinis und K. J. Bradley (2015, S. 161–164), dass die Performance in verschiedenen Berufs- und Hierarchieebenen eher einer Potenzgesetzverteilung

folgt, wie sie in der grauen Kurve in Abb. 4 dargestellt ist. Die Normalverteilung unterstellt, dass die meisten Performanceniveaus um einen Mittelwert μ_1 gruppiert sind, wobei sich sowohl hohe als auch niedrige Extremwerte symmetrisch verteilen (vgl. H. Aguinis und K. J. Bradley 2015, S. 162). Im Gegensatz dazu folgt die Potenzgesetzverteilung der Annahme, dass nur wenige Mitarbeitende hochgradig effektiv sind, während der Großteil eine unterdurchschnittliche Performance erbringt. Hierbei verschiebt sich der Mittelwert μ_2 im Vergleich zur Normalverteilung nach rechts (vgl. H. Aguinis und K. J. Bradley 2015, S. 161f.; K. R. Murphy 2020, S. 17f.). Unter der Potenzgesetzverteilung sind Maßnahmen, die die wenigen Leistungsträger demotivieren oder zu Genügsamkeit verleiten, deutlich kritischer als in der Normalverteilung. Diese Performer tragen überproportional zur Gesamtperformance bei (vgl. K. R. Murphy 2020, S. 17). Gleichzeitig stellt die Aktivierung des Leistungspotenzials der Majorität der eher leistungsschwachen Mitarbeitenden, die in Richtung des Durchschnitts entwickelt werden sollen, einen erheblichen Hebel für Produktivitätssteigerungen dar. In den betrachteten Fällen wird diese Gruppe jedoch oft demotiviert oder abgestoßen, während der Fokus auf den Top-Performern und dem gehobenen Mittelfeld liegt (vgl. H. Aguinis und K. J. Bradley 2015, S. 161f.; R. J. Akerlof und R. T. Holden 2012, S. 302–305; E. Dechenaux et al. 2015, S. 622–630; K. R. Murphy 2020, S. 16–19). Besonders ausgeprägt sind die Demotivationseffekte, wenn der Abstand zwischen den Performancerändern groß ist (vgl. P. Casas-Arce und A. Martínez-Jerez 2009, S. 1306; H. Eyring und G. Narayanan 2018, S. 606). Unter der Annahme einer Potenzgesetzverteilung erhält das U-förmige Reaktionsmuster auf Ranginformationen eine neue Bedeutung. Nicht nur die Sympathie mit der Ranglistenspitze sollte aktiviert werden, sondern auch die intrinsische Abneigung gegenüber einer schlechten Platzierung (vgl. D. Gill et al. 2019, S. 494f.).

Menschen reagieren je nach individueller Präferenz und Rangposition signifikant unterschiedlich, was dem prognostizierten Reaktionsmuster in klassischen Turniermodellen teilweise widerspricht (vgl. E. Cardinaels und C. Feichter 2023, S. 161f.; D. Gill et al. 2019, S. 503).

Zudem zeigt die Forschung, dass die Wirksamkeit von FR stark von der jeweiligen Person sowie der spezifischen Aufgabe abhängt. Aus diesem Grund wird zunehmend betont, dass die Kompatibilität von Turniersystemen mit der Aufgabenstellung, den beteiligten Personen und dem Umfeld genauer untersucht werden muss, um gegebenenfalls geeignete Alternativen zu entwickeln (vgl. T. Baldwin et al. 2013, S. 360–363; E. Cardinaels und C. Feichter 2023, S. 18). Die ökonomische Turnierforschung fokussierte sich hauptsächlich auf Detailuntersuchungen und den Austausch von Oberflächenmerkmalen, während sozial-psychologische Aspekte weitgehend unbeachtet blieben (vgl. E. Dechenaux et al. 2015, S. 610; R. Frederickson 1992, S. 647). R. Frederickson 1992 erweiterte den ökonomischen Mechanismus gemeinsamer Unsicherheiten erstmals um sozial-psychologisch begründete Verhaltensauswirkungen gemäß der TSV (vgl. R. Frederickson 1992, S. 647f.). Dabei wird angenommen, dass ein hohes Maß an Unsicherheit eine ausreichende Ähnlichkeit unter den Agenten vermuten lässt, wodurch SV eine relevante Anreizwirkung entfaltet. Es wird anerkannt, dass sozialer Einfluss im RPM affektive Reaktionen auslöst, die das Verhalten sowohl positiv als auch negativ beeinflussen können. Über monetäre Anreize hinaus existieren demnach auch intrinsische, verhaltensrelevante Aspekte, die Mitarbeitende zu mehr Anstrengung und höherer Performance motivieren können (vgl. L. Festinger 1954, S. 117–120; R. Frederickson 1992, S. 648–654; S. Goebel und E. Weißenberger 2017, S. 192; S. Murthy und A. Schafer 2011, S. 160; C. Schnieder 2022, S. 75; A. Tesser 1988, S. 182). Aus dieser Perspektive sind Turniere und Ranglistensysteme nicht alternativlos. Zwar kann diese Form von RPM im Durchschnitt die Leistung steigern und behält seine Berechtigung, jedoch birgt es auch die Gefahr, gegenteilige Effekte hervorzurufen (vgl. D. Gill et al. 2019, S. 503). Um Anreizsysteme im MC nachhaltig zu verbessern, ist ein tieferes Verständnis der sozial-psychologischen Auswirkungen notwendig, die durch RPM-Systeme ausgelöst werden (vgl. L. Hannan et al. 2008, S. 894; F. Hartmann und P. Schreck 2018, S. 363f.).

4 Berücksichtigung sozial-psychologischer Komponenten bei der Anreizwirkung der relativen Messung

Die Literatur zu Anreizwirkungen der TSV im Kontext von RPM untersucht zunehmend Steuerungssysteme, bei denen RPM nicht direkt mit einer Vergütungskomponente verknüpft ist. Ziel ist es, Verhaltenseffekte auf Anstrengung und Performance isoliert zu analysieren und ein Verständnis jenseits der ökonomischen Anreizhypothese zu entwickeln (vgl. F. Hartmann und P. Schreck 2018, S. 1189; J. Luft 2016, S. 78; S. Murthy und A. Schafer 2011, S. 554; C. Schnieder 2022, S. 72–75; D. Tafkov 2013, S. 329). Um fundierte Aussagen zur verbesserten Gestaltung von Anreizsystemen in diesem Kontext zu ermöglichen, wird der in Abb. 5 dargestellte konzeptionelle Forschungsrahmen kurz eingeführt.

Die grundlegende Wirkungskette zeigt, dass RPM den SV-Prozess aktiviert und in der daraus entstehenden Wettbewerbsumgebung, abhängig von der Prozessrichtung, für das Unternehmen positive oder negative Verhaltenseffekte hervorrufen kann. Primär wird ein Anstrengungseffekt angenommen, der als zentraler Prädiktor auch einen Performanceeffekt bedingt (vgl. E. Bonner und B. Sprinkle 2002, S. 304–306; J. Luft 2016, S. 78; C. Schnieder 2022, S. 76). Anstrengung lässt sich in Richtung, Dauer und Intensität differenzieren. Die Richtung gibt an, in welche Aufgabe oder Tätigkeit ein Mitarbeitender seine Energie investiert, die Dauer beschreibt die Zeitspanne des Ressourceneinsatzes, die Intensität misst die tatsächliche Aufmerksamkeit innerhalb dieser Zeitspanne (vgl. G. Birnberg et al. 2006, S. 117; E. Bonner und B. Sprinkle 2002, S. 304–306). Der konzeptionelle Rahmen dieses Kapitels wird durch drei Variablenkategorien erweitert, die neben den Bereitstellungsparametern des RPM-Systems (A) den SV-Prozess

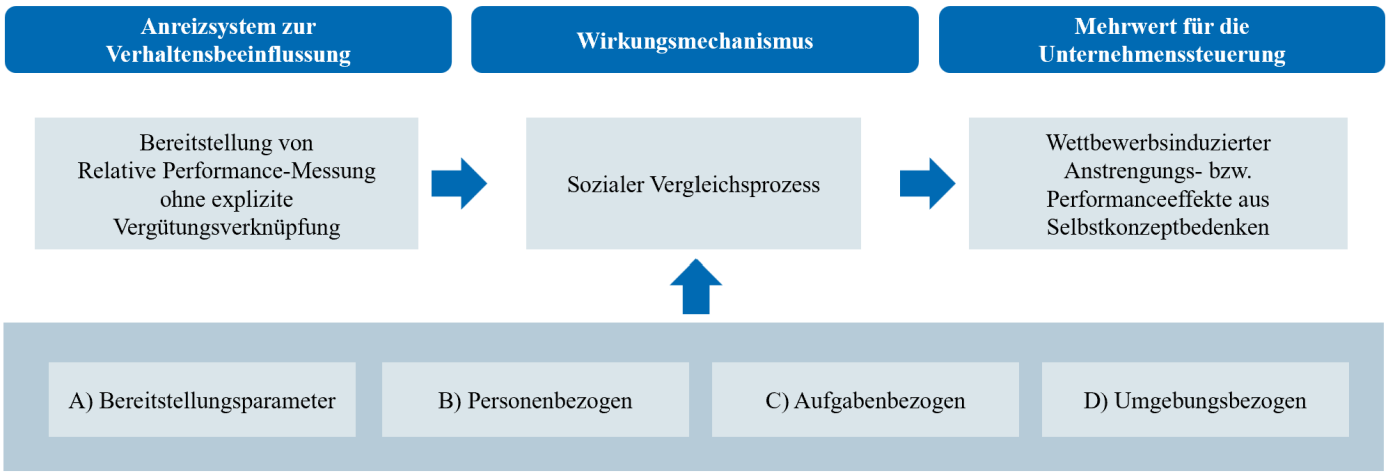

Abbildung 5: Rahmen einer auf der Theorie des sozialen Vergleichs konzentrierten Forschung.

(Quelle: vgl. E. Bonner und B. Sprinkle 2002, S. 312–326; M. Garcia et al. 2013, S. 634–637; C. Schnieder 2022, S. 74–78.)

maßgeblich moderieren. Diese umfassen personenbezogene (B), aufgabenbezogene (C) und umgebungsbezogene (D) Variablen (vgl. E. Bonner 1999, S. 390; E. Bonner und B. Sprinkle 2002, S. 310–326; M. Garcia et al. 2013, S. 634–637; J. Luft 2016, S. 78; C. Schnieder 2022, S. 77f.). Ihre Relevanz zeigt sich sowohl in der allgemeinen Analyse der Wirksamkeit von Anreizsystemen (vgl. E. Bonner 1999, S. 390; E. Bonner und B. Sprinkle 2002, S. 312–326), in der Erklärung von Wettbewerbsverhalten unter dem Einfluss des SV-Prozesses (vgl. M. Garcia et al. 2013, S. 634–637) als auch in der spezifischen Untersuchung von isoliertem RPM, wie in Abb. 5 dargestellt (vgl. C. Schnieder 2022, S. 74–78).

In Anlehnung an E. Bonner und B. Sprinkle 2002 lässt sich die Performance wie in Formel (5) darstellen:

$$Performance = f_e(A, B, C, D) \tag{5}$$

Die Performancefunktion ist in diesem Fall durch Anstrengung vermittelt und abhängig von den Variablen A-D (vgl. E. Bonner und B. Sprinkle 2002, S. 304, 311; D. Tafkov 2013, S. 330). Die Bereitstellungsparameter bestimmen die zentralen Attribute des RPM-Systems, darunter die Wahl der Form (siehe Abb. 1), die damit verbundene Genauigkeit sowie den Öffentlichkeitsgrad. Sie legen somit die Konfiguration des RPM fest (vgl. L. Hannan et al. 2019, S. 1608f.; C. Schnieder 2022, S. 77; D. Tafkov 2013, S. 330–333). Personenbezogene Variablen beschreiben individuelle Eigenschaften wie Persönlichkeit, Wertesystem, Wettbewerbspräferenz und die Fähigkeit, eine Aufgabe zu bewältigen. Diese Variablen prägen den Grundzustand des Selbst (vgl. E. Bonner und B. Sprinkle 2002; N. Eichhorn 2016, S. 106–308; M. Garcia et al. 2013, S. 636f.; C. Schnieder 2022, S. 78). Aufgabenbezogene Variablen umfassen die Komplexität und Dimensionalität der Aufgabe, die durch das RPM adressiert wird (vgl. E. Bonner und B. Sprinkle 2002, S. 319; N. Eichhorn 2016, S. 100f.; L. Gerhards und N. Siemer 2016, S. 1189; C. Schnieder 2022, S. 78). Umgebungsbezogene Variablen charakterisieren den Kontext, in dem die Aufgabe ausgeführt wird. Besonders relevant ist hierbei das hintergründige, vom RPM unabhängige Vergütungssystem (vgl. E. Bonner und B. Sprinkle 2002, S. 325; M. Garcia et al. 2013, S. 637; J. Luft 2016, S. 78; C. Schnieder 2022, S. 78).

Zahlreiche Belege zeigen, dass RPM die Anstrengung und Performance beeinflusst. Die Schwankungen in der Richtung dieser Effekte verdeutlichen jedoch die Notwendigkeit, die in Abb. 5 abgeleiteten Variablenkategorien genauer zu untersuchen (vgl. C. Schnieder 2022, S. 83–92). Mitarbeitende erhalten im Rahmen ihrer Beschäftigung stets eine Entlohnung. Aus der ökonomischen Hypothese ist bekannt, dass monetäre Anreize Einfluss auf Motivation und Verhalten haben (vgl. C. Schnieder 2022, S. 87). Studien haben gezeigt, dass das Vergütungssystem als Variable auch dann eine Rolle spielt und kontrolliert werden sollte, wenn RPM nicht unmittelbar mit einer finanziellen Komponente verknüpft ist. Dabei wird grundlegend zwischen Pauschalsystemen (fix) und Stückvergütungssystemen (outputbasiert) unterschieden (vgl. L. Berger et al. 2019, S. 20; S. Blader et al. 2019, S. 39; N. Eichhorn 2016, S. 110–114; L. Hannan et al. 2013, S. 554; S. Murthy 2010, S. 1–3; S. Murthy und A. Schafer 2011, S. 161f.; D. Tafkov 2013, S. 328–330; A. Tran und R. Zeckhauser 2012, S. 645f.). S. Murthy 2010 untersucht in einem Laborexperiment mit mehreren Perioden und realer Anstrengung, ob RPM die Motivation der Teilnehmenden steigern kann und wie sich diese Wirkung zwischen Stücklohnvergütung mit konstanter Rate sowie quoten- und budgetbasierter Vergütung unterscheidet. Quo-

tenbasierte Vergütung umfasst eine Pauschalvergütung, die ab einer Zielmarke durch einen Stückbonus ergänzt wird, während bei budgetbasierter Vergütung eine Stückzahlung ab der Zielmarke ansteigt (vgl. S. Murthy 2010, S. 1–3). Die Ergebnisse zeigen unter allen Vergütungsbedingungen positive Anstrengungseffekte im Vergleich zur Situation ohne RPM-Behandlung, wobei die stärkste Steigerung bei der quotenbasierten Vergütung erzielt wird (vgl. S. Murthy 2010, S. 3f.). Weitere Studien bestätigen einen durchschnittlich positiven Effekt verschiedener RPM-Systeme unter Anwesenheit von Pauschal- und Stückvergütung (vgl. L. Berger et al. 2019, S. 20; L. Hannan et al. 2013, S. 555; L. Hannan et al. 2008, S. 903–905; D. Tafkov 2013, S. 328f.; W. Wang 2017, S. 968).

D. Tafkov (2013) (S. 327–332, 347) belegt, dass der wettbewerbsinduzierte Effekt von RPM in Übereinstimmung mit der ökonomischen Theorie durch hintergründige Stückvergütung stärker ausgeprägt ist als unter Pauschalvergütung. Er untersucht ebenfalls den Vergleich zwischen privatem und öffentlichem RPM. Im privaten RPM erhält der Einzelne ausschließlich seine eigene Ranginformation, wobei die Anonymität gewahrt bleibt. Im öffentlichen RPM hingegen werden die Verteilungsinformationen aller Teilnehmenden neben der eigenen angezeigt (vgl. D. Tafkov 2013, S. 327f.). Öffentlichkeit spielt also eine wichtige Rolle im SV-Prozess. Das Individuum zeigt bei öffentlicher Information ein höheres Involvement, da die Relevanz des Vergleichs steigt. Der soziale Druck und die potenzielle Bedrohung des Selbstwertes nehmen zu, während beide emotionalen Pole, Scham und Anerkennung, stärker wirken (vgl. D. Tafkov 2013, S. 328–333). Der Effekt der Öffentlichkeit wird auch in einer Untersuchung von A. Tran und R. Zeckhauser (2012, S. 646f.) bestätigt. Teilnehmer eines Semesterkurses wurden in drei Gruppen unterteilt, eine ohne RPM-Behandlung innerhalb der Belegdauer, eine mit privatem und eine mit öffentlichem RPM. Die Abschlussergebnisse, bei denen maximal 700 Punkte erreichbar sind, verdeutlichen den RPM-Einfluss. Die private Gruppe erzielt 52 Punkte mehr als die Kontrollgruppe. Die öffentliche Gruppe übertrifft sowohl die Kontrollgruppe um 76 Punkte als auch die private Gruppe um 24 Punkte. Dies zeigt einen signifikanten Effekt der RPM-Behandlung, der durch öffentliche Informationen weiter verstärkt wird (vgl. A. Tran und R. Zeckhauser 2012, S. 648). Die öffentliche Bekanntgabe der RPI steigert die Wettbewerbsaktivität der Teilnehmenden. Zudem weisen starke Hinweise darauf hin, dass die wettbewerbsfördernden Effekte von Stückvergütung und erhöhtem Öffentlichkeitsgrad kumulativ wirken (vgl. S. Blader et al. 2019, S. 722–729; L. Gerhards und N. Siemer 2016, S. 1188f.; L. Hannan et al. 2013, S. 555f.; H. Song et al. 2018, S. 2628f.; D. Tafkov 2013, S. 328–333; A. Tran und R. Zeckhauser 2012, S. 646–650; W. Wang 2017, S. 968). In einer Feldstudie, die in zwei Notaufnahmen unter Ärzten durchgeführt wurde, konnte der Einfluss des Öffentlichkeitsgrads in einem Setting untersucht werden, in dem die Mitarbeitenden die Maßnahmen der komplexen Patientenbehandlung selbst wählen. Zunächst unterliegen beide Notaufnahmen privatem RPM über den Behandlungserfolg, wobei eine der Gruppen später auf öffentliches RPM umstellt (vgl. H. Song et al. 2018, S. 2628, 2636). Nach dem Wechsel verzeichnen H. Song et al. (2018, S. 2628f.) eine durchschnittliche Produktivitätssteigerung von 10,9 %. Dies wird darauf zurückgeführt, dass öffentliches RPM nicht nur die Identifikation von Leistungsträgern erleichtert, sondern auch den Austausch bewährter Verfahren durch deren Erfolg fördert. Besonders leistungsschwache Ärzte verzeichnen eine zusätzliche Produktivitätssteigerung von 8,9 %, um sich den Top-Performern anzupassen, was die Performancedifferenz zwischen den Teilnehmenden verringert und die Wahl geeigneter Aufgabenstrategien zur Selbstverbesserung unterstützt (vgl. H. Song et al. 2018, S. 2629–2632).

Andere Autoren untersuchen die Unterscheidung zwischen einer auf Anstrengung basierenden Aufgabe sowie einer intelligenz- und kreativitätsorientierten Aufgabe, bei der Prestige stärker aus zentralen Fähigkeiten resultiert. Die Ergebnisse, die unter Pauschalvergütung erfasst wurden, zeigen eine robuste Wirkung sowohl unter privatem als auch öffentlichem RPM (vgl. L. Gerhards und N. Siemer 2016, S. 1189). Es wird festgestellt, dass eine einsatzorientierte Aufgabe gut auf den Wettbewerbsmechanismus anspricht und einen Anstrengungseffekt über die gesamte Performanceverteilung hinweg erzielen kann. Hingegen scheint RPM bei Aufgaben, die stärker auf individuelle Fähigkeiten angewiesen sind, insb. bei leistungsschwächeren Individuen, negative Emotionen zu verstärken. Folgen sind der Verlust an Selbstwertgefühl, dysfunktionale Aufgabenstrategien und reduzierte Anstrengung (vgl. L. Gerhards und N. Siemer 2016, S. 1194–1199). Eine weitere Facette der Untersuchungen bezieht sich auf mehrdimensionale Aufgaben, die entweder mehrere Teilaufgaben kombinieren oder eine Gesamtaufgabe mit Verantwortung für sowohl Qualität als auch Quantität umfassen. Beispiele hierfür sind komplexe Produktionsaufgaben sowie die Tätigkeiten von Ärzten oder Managern (vgl. N. Eichhorn 2016, S. 101; L. Hannan et al. 2013, S. 554). In solchen Szenarien stellt sich nicht nur die Frage nach der Anstrengung des Mitarbeitenden, sondern auch, wie sich diese Anstrengung auf die verschiedenen Dimensionen einer Aufgabe verteilt. RPM kann daher zwei Effekte hervorrufen: einen Motivationseffekt zur Steigerung der Gesamtanstrengung und einen Anstrengungsverzerrungseffekt (vgl. N. Eichhorn 2016, S. 100f.; L. Hannan et al. 2013, S. 553f.). Aus den Unternehmenszielen kann eine gewünschte Allokation entlang der Dimensionen abgeleitet werden. Eine Anstrengungsverzerrung beschreibt den Fall, wenn die tatsächliche Verteilung der Gesamtanstrengung von der gewünschten abweicht (vgl. L. Hannan et al. 2013, S. 554). Wird RPM über mehrere Dimensionen einer Aufgabe bereitgestellt, besteht die Gefahr, dass Mitarbeitende unverhältnismäßig viel Anstrengung auf eine einzelne Dimension konzentrieren. Dieses Verhalten der einseitigen Positionsverbesserung zielt darauf ab, eine Selbstwertbedrohung in einer anderen Dimension zu kompensieren (vgl. N. Eichhorn 2016, S. 112, 128f.; L. Hannan et al. 2013, S. 555–558). Diese verzerrte Anstrengungsverteilung kann bei Aufgaben mit abnehmendem Grenzertrag die Gesamtperformance

negativ beeinflussen, obwohl die Gesamtanstrengung steigt (vgl. L. Berger et al. 2019, S. 1610; N. Eichhorn 2016, S. 117–119; L. Hannan et al. 2013, S. 555f.). Bei mehrdimensionalen Aufgaben tritt eine Anstrengungsverzerrung jedoch erst durch die Aktivierung des SV-Prozesses auf, wobei öffentliches RPM diesen Effekt zusätzlich verstärkt (vgl. L. Hannan et al. 2013, S. 555–558). Eine alternative Herangehensweise empfiehlt die Nutzung eines eindimensionalen RPM-Systems bei mehrdimensionalen Aufgaben. Dieses wählt ein aggregiertes Performancemaß, das alle Dimensionen vereint und eine Gesamtbewertung der Fähigkeiten liefert, anstatt separate Bewertungen für jede einzelne Dimension zu ermöglichen (vgl. N. Eichhorn 2016, S. 111f.).

L. Hannan et al. 2019 vertiefen die Analyse der Anstrengungsverzerrung und untersuchen die Auswirkungen zweier Designkomponenten auf diesen dysfunktionalen Effekt (vgl. L. Hannan et al. 2019, S. 1607). Diese Designkomponenten umfassen die Detailebene und die zeitliche Aggregation des RPM. Bei der Detailebene wird zwischen einer rein ordinalen Rangfolge und der zusätzlichen Angabe kardinaler Ergebnisse differenziert. Die zeitliche Aggregation variiert zwischen Ein-Perioden-RPM und aggregierten Ergebnissen über einen Gesamtzeitraum (vgl. L. Hannan et al. 2019, S. 1607–1609). Die Aussagekraft des RPM, definiert als Grad der Klarheit über die relative Performanceposition, steigt bei kardinalen und aggregierten Informationen. Mit zunehmender Aussagekraft nimmt jedoch auch die Ausprägung der Anstrengungsverzerrung zu. In diesem Zusammenhang erscheint ein kardinales und aggregiertes RPM-Design sowohl einzeln als auch kombiniert kritischer als ein ordinales und zurückgesetztes RPM (vgl. L. Hannan et al. 2019, S. 1608–1610).

Abgesehen von der Dimensionalität der Aufgabe und des RPM spielt die Darstellung des RPM-Designs eine wichtige Rolle. Dabei zeigen sich insb. unterschiedliche Reaktionsmuster zwischen leistungsstarken und leistungsschwachen Gruppen (vgl. H. Eyring und G. Narayanan 2018, S. 581–584; S. Kramer et al. 2016; S. Murthy und A. Schafer 2011, S. 160f., 17f.). S. Murthy und A. Schafer (2011, S. 160) vergleichen in ihrer Untersuchung die Auswirkungen ohne RPM mit denen einer ordinalen und einer prozentualen Darstellung. Unabhängig von der Performanceposition wird die Information zudem mit einem positiven, neutralen oder negativen Framing versehen. Ein solches Framing beeinflusst und intensiviert potenziell die affektive Reaktion (vgl. S. Murthy und A. Schafer 2011, S. 163). Im Durchschnitt verzeichnen beide RPM-Formen unter allen Framing-Varianten im Vergleich zur Abwesenheit von RPM eine positive Wirkung auf die Performance (vgl. S. Murthy und A. Schafer 2011, S. 160f.). Leistungsschwache zeigen unter positivem Framing eine signifikante Leistungssteigerung, da trotz schlechter Position Hoffnung und ein Gefühl erreichbarer Selbstverbesserung geweckt werden. Negatives Framing hingegen verstärkt Gefühle von Scham, führt zu Entmutigung und senkt die Performance sogar unter das Niveau neutraler Informationen. Leistungsstarke reagieren in beiden Framing-Szenarien mit einer erhöhten Performance gegenüber neutralem RPM, zeigen jedoch unter negativem eine stärkere Steigerung. Dies deutet darauf hin, dass die Sorge, ihre vorteilhafte Position zu verlieren, für diese Gruppe motivierender ist als positive Selbstbestätigung (vgl. S. Murthy und A. Schafer 2011, S. 180f.). S. Kramer et al. (2016, S. 161f.) untersuchen neben dem Framing auch die Wirkung unterschiedlicher Sortierungen einer Rangfolge. Dabei wird zwischen einer ungeordneten Rangliste, einer Sortierung vom Besten zum Schlechtesten (Gewinnerrangliste) und vom Schlechtesten zum Besten (Verliererrangliste) variiert. Im Durchschnitt führt RPM zu einem positiven Performanceeffekt (vgl. S. Kramer et al. 2016, S. 19). Während bei Leistungsstarken die Sortierung keine signifikanten Unterschiede in der Reaktion bewirkt, zeigen Leistungsschwache unter der Verliererrangliste eine deutlich stärkere Verbesserung. Die Hervorhebung des absoluten Verlierers erzeugt zusätzlichen sozialen Druck. Das Ziel, Scham durch eine Platzierung an der Spitze der Verliererliste zu vermeiden, intensiviert den SV-Prozess (vgl. S. Kramer et al. 2016, S. 16–19). Abweichend von der gut untersuchten Präsentation als Rangfolge (vgl. W. Chan 2018, S. 84; L. Hannan et al. 2013, S. 554f.; S. Murthy und A. Schafer 2011, S. 159; C. Schnieder 2022, S. 84–87; D. Tafkov 2013, S. 328–333) schlagen H. Eyring und G. Narayanan (2018, S. 581) in einem Feldexperiment die Wahl eines Referenzpunkts zum Vergleich vor. In ihrer Studie wird die Wirkung des Medians und einer spezifischen Stelle im obersten Quantil als Referenzpunkt im RPM-System verglichen. Es zeigt sich, dass sowohl die Wahl des Referenzpunktes als auch die Position des Einzelnen in der Performanceverteilung eine entscheidende Rolle für die affektive Reaktion und den Verhaltenseffekt im SV-Prozess spielen (vgl. H. Eyring und G. Narayanan 2018, S. 581–583). Der Median gilt für viele leistungsschwache Individuen als erreichbar, was eine Selbstverbesserung aktivieren kann. Hingegen löst der Referenzpunkt des oberen Quantils aufgrund seiner Unerreichbarkeit Schamgefühle aus, was zu einem negativen Performanceeffekt führt. Für leistungsstarke Individuen wird lediglich bei einem höheren Referenzpunkt ein positiver Leistungseffekt erzielt, während der Median an der Leistungsspitze keine motivierenden Emotionen hervorruft (vgl. H. Eyring und G. Narayanan 2018, S. 582–585). Im Durchschnitt haben weder der Median noch das obere Quantil einen positiven Performanceeffekt, was nahelegt, dass der Referenzpunkt im RPM-System je nach Leistungsgruppe angepasst werden sollte (vgl. H. Eyring und G. Narayanan 2018, S. 606f.).

Die Literatur liefert Hinweise darauf, dass die Reaktion auf RPM-Systeme auch von individuellen Persönlichkeitsmerkmalen abhängt, wobei der Wirkungsmechanismus vor allem auf einer wettbewerbsinduzierten Verhaltensanpassung basiert (vgl. W. Wang 2017, S. 966f.). F. Thomas und A. Thornock (2015) (S. 26f.) unterscheiden in diesem Kontext zwischen Pro-Selbst- und Pro-Sozialen Individuen. Pro-Selbst bezeichnet eigennützige, egozentrische Menschen, die von Natur aus wettbewerbsorientiert entscheiden und handeln. Diese Personen tendieren dazu,

in einen SV-Prozess einzutreten, was den Mechanismen der TSV entspricht. Im Gegensatz dazu neigen Pro-Soziale Individuen zu altruistischem Verhalten und einem Fokus auf das Gemeinwohl. Unter RPM könnten unerwartet dysfunktionale Verhaltensweisen auftreten. Beispielsweise könnten leistungsstarke Personen absichtlich ihre Performance senken, um den Selbstwert der schwächeren Mitbewerber zu schützen (vgl. S. Blader et al. 2019, S. 743; F. Thomas und A. Thornock 2015, S. 26–29). W. Wang 2017 berichtet ähnliche Ergebnisse und untersucht die Reaktion auf RPM in Abhängigkeit vom Grad der dunklen Triade. Dieses Persönlichkeitskonstrukt, bestehend aus Narzissmus, Machiavellismus und Psychopathie, weist eine starke Ähnlichkeit zur Definition des Pro-Selbst-Charakters auf (vgl. W. Wang 2017, S. 967). Pro-Selbst-Individuen neigen dazu, sich stärker mit dem Kompetenzwettbewerb in Fähigkeiten auseinanderzusetzen, was unter der Präsenz von RPM eher zu einem aufgabenorientierten Wettbewerb führt und einen produktiven Anstrengungszuwachs fördert. Im Gegensatz dazu reagieren Individuen mit einem geringen Grad an dunkler Triade unterschiedlich auf RPM. Diese Personen tendieren weniger zum Vergleich in Fähigkeiten und neigen stattdessen intuitiv zu einem rivalitätsorientierten Wettbewerb, was zu einem kontraproduktiven Anstrengungszuwachs und dysfunktionalem Verhalten führen kann. Ein Verständnis für die persönlichkeitsgetriebenen Reaktionsmuster ist für Unternehmen relevant, da anzunehmen ist, dass je nach Bereich tendenziell Mitarbeitende mit ähnlichen Persönlichkeitsmerkmalen angezogen werden (vgl. F. Thomas und A. Thornock 2015, S. 26–29; W. Wang 2017, S. 967–969).

5 Implikationen für die Gestaltung effizienter Anreizsysteme

Für Unternehmen stellt RPM eine Option zur Unternehmenssteuerung mit geringen direkten Kosten dar (vgl. J. Luft 2016, S. 76f.; A. Merchant und Van der Stede 2017, S. 174; D. Tafkov 2013, S. 329; W. Wang 2017, S. 967). Zudem basiert RPM auf einem intuitiven Wirkungsmechanismus, der in der menschlichen Natur verankert ist und über den Arbeitskontext hinaus Relevanz besitzt (vgl. H. Beach und A. Tesser 1995, S. 146–148; L. Festinger 1954, S. 118–123; D. Tafkov 2013, S. 330f.; A. Tesser 1988, S. 204–206). Die Einführung von RPM erfordert jedoch eine sorgfältige Planung und ist kein Selbstzweck. Einerseits muss der SV-Prozess in ausreichendem Maße aktiviert werden, um die gewünschten Effekte zu erzielen. Andererseits sind die potenziellen indirekten Kosten, die aus unerwünschten affektiven Reaktionen und dysfunktionalen Verhaltensstrategien im Wettbewerb resultieren können, zu berücksichtigen. Eine zielgerichtete Gestaltung des Anreizsystems ist daher essentiell (vgl. L. Berger et al. 2019, S. 19f.; N. Eichhorn 2016, S. 126; D. Tafkov 2013, S. 329; W. Wang 2017, S. 969). Abb. 6 bietet eine Übersicht über sechs zentrale Gestaltungskriterien, die zur Bewältigung der Herausforderungen bei der Planung des MC beitragen.

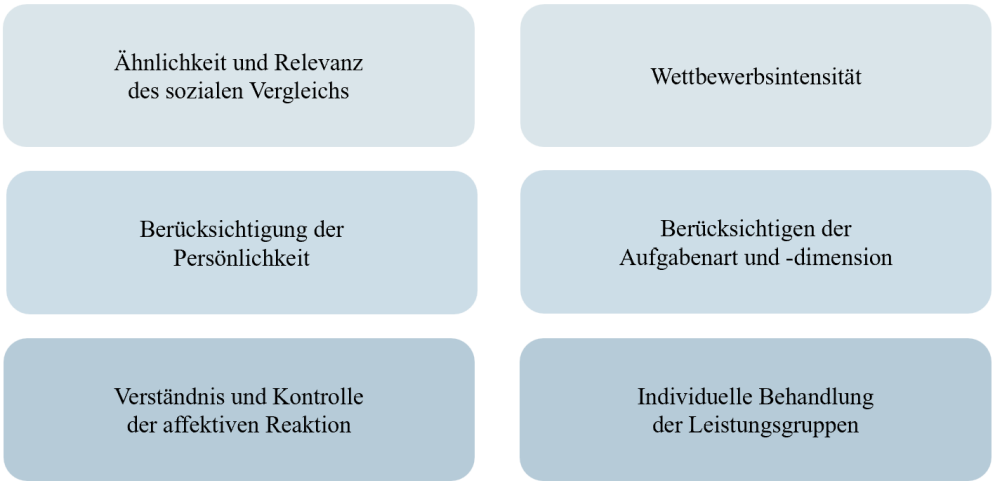

Abbildung 6: Gestaltungskriterien für relative Performance-Messung.

Die TSV verdeutlicht, dass Mitarbeitende, die durch RPM gesteuert werden sollen, ein Mindestmaß an Ähnlichkeit aufweisen müssen, damit der SV in der jeweiligen Arbeitssituation ausreichend relevant ist. Dies ermöglicht die Schaffung einer Wettbewerbsumgebung, in der potenzielle Motivations-, Anstrengungs- und Performanceeffekte aktiviert werden können (vgl. D. Tafkov 2013, S. 330–332). Im Einklang mit der AT lässt sich dies auch als das kritische Maß gemeinsamer Unsicherheiten in den zu bewältigenden Aufgaben definieren (vgl. T. Ball et al. 2020, S. 1510; A. Dye 1992, S. 45; R. Frederickson 1992, S. 650–653; D. Tafkov 2013, S. 330f.).

Die durch RPM ausgelösten Prozesse interagieren empfindlich mit dem bestehenden Vergütungssystem, selbst wenn das Steuerungsinstrument nicht unmittelbar daran gekoppelt ist. Die Nützlichkeit von RPM muss daher stets im Kontext des gewählten Vergütungsmodells bewertet

werden (vgl. L. Hannan et al. 2008, S. 897, 910). Eine outputbasierte Vergütung intensiviert den Wettbewerb durch ihren ökonomisch-monetären Anreiz (vgl. E. Bonner und B. Sprinkle 2002, S. 19f.; L. Hannan et al. 2008, S. 895; J. Luft 2016, S. 78; C. Schnieder 2022, S. 72f.). Ebenso verstärkt ein höherer Öffentlichkeitsgrad von RPM den Wettbewerb, da soziale Informationen kritischer auf den Selbstwert wirken (vgl. J. Luft 2016, S. 78; C. Schnieder 2022, S. 72f.; D. Tafkov 2013, S. 347). Ein intensiverer Wettbewerb führt jedoch nicht zwangsläufig zu verbesserten Anstrengungs- und Performanceeffekten. Die Effektrichtung hängt entscheidend von der Art des Wettbewerbs und der individuellen affektiven Reaktion auf das geschaffene Umfeld ab. Ein durch SV initiierter Wettbewerb kann als Effizienzmotor wirken, jedoch auch rivalisierendes Denken fördern, was Anstrengung in dysfunktionales Verhalten oder suboptimale Aufgabenstrategien umleiten kann (vgl. H. Song et al. 2018, S. 2644; D. Tafkov 2013, S. 347; A. Tran und R. Zeckhauser 2012, S. 649). Je stärker eine Persönlichkeit der Pro-Selbst-Definition entspricht, desto wahrscheinlicher ist ein wettbewerbsorientiertes Verhalten (vgl. F. Thomas und A. Thornock 2015, S. 27). Diese Charaktere zeigen eine Neigung zu fähigkeitsorientierter Konkurrenz, bei der eher ein konstruktiver Anstrengungszuwachs zu erwarten ist (vgl. W. Wang 2017, S. 969). Im Gegensatz dazu reagieren gemeinwohlorientierte, pro-soziale Persönlichkeiten schwächer auf Wettbewerb, schützen eher die Emotionen anderer oder entwickeln kontraproduktive Anstrengungsmuster durch rivalitätsorientierte Wettbewerbsantworten (vgl. F. Thomas und A. Thornock 2015, S. 27; W. Wang 2017, S. 969). Unternehmen sollten den Einsatz von RPM daher sorgfältig im Hinblick auf die vorherrschenden Persönlichkeitsprofile und die gelebten Werte prüfen. Insbesondere in sozialen Branchen und Arbeitsfeldern könnte Zurückhaltung geboten sein (vgl. W. Chan 2018, S. 969f.; F. Thomas und A. Thornock 2015, S. 27–29).

Die Aufgabenart sollte so gewählt sein, dass sie einen effizienzsteigernden Wettbewerb ermöglicht, ohne die Aufmerksamkeit und Konzentration von der eigentlichen Aufgabenstrategie abzulenken. Dabei ist insbesondere zwischen einfachen, einsatzorientierten Aufgaben und solchen mit höherer Komplexität und Dimensionalität zu unterscheiden (vgl. L. Gerhards und N. Siemer 2016, S. 1194–1199; H. Song et al. 2018, S. 2644). In einem Multi-Tasking-Setting und beim Einsatz von RPM über mehrere Dimensionen sollte die Implementierung davon abhängen, ob das Unternehmen die Allokation des Gesamtaufwands kontrollieren und steuern kann. Um dysfunktionale Anstrengungsverzerrungen zu vermeiden, sollte RPM darauf abzielen, Selbstwertbedrohungen durch die Betonung einzelner Dimensionen zu minimieren und stattdessen die Gesamtanstrengung zu fördern (vgl. N. Eichhorn 2016; L. Hannan et al. 2019, S. 1610; L. Hannan et al. 2013, S. 554–556). Eine Überwachung der Anstrengungsverteilung, einschließlich Unter- und Obergrenzen für jede Dimension kann positive Effekte auf die Gesamtperformance sicherstellen (vgl. L. Hannan et al. 2019, S. 1608–1610; L. Hannan et al. 2013, S. 554–556). Dennoch birgt der Einsatz von mehrdimensionalem RPM Risiken für die Unternehmensziele. Solche Maßnahmen machen zusätzliche Kontroll- und Überwachungsmechanismen erforderlich, die mit höheren direkten Kosten verbunden sind (vgl. N. Eichhorn 2016, S. 133; L. Hannan et al. 2013, S. 554–556). Der Einsatz eines aggregierten Gesamtperformancemaßes über alle Dimensionen hinweg stellt eine effizientere Methode dar, die Ziele des MC bei Mehrdimensionalität besser zu erfüllen. Diese simulierte Eindimensionalität fokussiert den Wettbewerb auf die Gesamtaufgabenperformance und verhindert weitgehend Anstrengungsverzerrungen (vgl. N. Eichhorn 2016, S. 126–129).

Die Potenzgesetzverteilung der Performance unter Mitarbeitenden hebt zwei zentrale Gruppen für das MC hervor. Zum einen die wenigen überproportionalen Top-Performer, deren Leistungsniveau durch ein RPM-System stabilisiert und vor übermäßiger Selbstzufriedenheit geschützt werden sollte. Zum anderen die große Masse der Leistungsschwachen, die ein erhebliches Potenzial zur signifikanten Steigerung der Unternehmensperformance in Richtung Median aufweist (vgl. H. Aguinis und K. J. Bradley 2015, S. 161f.; K. R. Murphy 2020, S. 17–20). Studien zeigen, dass gerade bei Leistungsschwachen der SV-Prozess von besonderer Bedeutung ist. Hier treten Selbstwertbedenken, die Tendenz zur Assimilation nach oben und Selbstverbesserung besonders ausgeprägt auf. Allerdings ist Vorsicht geboten. Überschreitet die emotionale Belastung ein kritisches Maß, können Demotivation, Anstrengungsreduktion oder gar ein vollständiges Aufgeben die Folge sein (vgl. Y. Chen et al. 2010, S. 1374; S. Kramer et al. 2016, S. 17–49; S. Murthy und A. Schafer 2011, S. 161; H. Song et al. 2018, S. 2644; D. Tafkov 2013, S. 437; A. Tran und R. Zeckhauser 2012, S. 649). Ranglisten können durch eine geschickte Sortierung so gestaltet werden, dass das Wettbewerbsziel für Leistungsschwache darin besteht, nicht der absolute Verlierer zu sein. Dies verhindert eine Entmutigung durch als unerreichbar empfundene Spitzenleistungen (vgl. L. Hannan et al. 2008, S. 897; S. Kramer et al. 2016, S. 21–23). Ein positives Framing relativ schwacher Leistungen kann die affektive Reaktion in Richtung Selbstverbesserung lenken. Gleichzeitig zeigt ein negatives Framing bei Top-Performern durch kontrastive Vergleiche und die Sorge, ihre vorteilhafte Position zu verlieren, einen verstärkten Motivationseffekt (vgl. E. Carroll und D. Marginson 2021, S. 13; S. Murthy und A. Schafer 2011, S. 181). Durch die Nutzung individueller, an die Performancegruppe angepasster Referenzpunkte innerhalb der Gesamtverteilung, anstelle einer starren Rangliste, ist eine gezielt ausgewogene Incentivierung möglich. Indem der SV jeweils auf eine fokussierte Spannbreite begrenzt wird, sinkt das Risiko der Selbstgefälligkeit bei Leistungsstarken, während ein realistisches Ziel für Leistungsschwache erhalten bleibt. Der Referenzpunkt sollte dynamisch an das Performanceniveau angepasst werden (vgl. H. Eyring und G. Narayanan 2018, S. 586–606; L. Hannan et al. 2008, S. 897; H. Song et al. 2018, S. 2644). Die Gestaltung eines RPM-basierten Anreizsystems muss

die sechs zentralen Kriterien im Kontext der komplexen Wechselwirkung zwischen Umfeld, Aufgaben und Mitarbeitenden berücksichtigen. Nur so kann das Potenzial von RPM im individuellen Anwendungsfall genutzt werden, um gewünschte Performanceeffekte zu fördern und gleichzeitig Dysfunktionalitäten und indirekte Kosten zu minimieren.

6 Fazit und Ausblick

Obwohl RPM-Systeme weit verbreitet eingesetzt werden, bleibt die Schwierigkeit, sie so zu gestalten, dass sowohl die langfristige Motivation der Mitarbeitenden als auch die Strategie des Unternehmens gefördert werden. Das übergeordnete Ziel des MC ist die Harmonisierung des unternehmerisch gewünschten mit dem tatsächlichen Handeln der Mitarbeitenden. Vor diesem Hintergrund ergibt sich die Leitfrage nach den Herausforderungen bei der Gestaltung von Anreizsystemen auf Basis von RPM, um durch gezielte Verhaltensbeeinflussung die Performance zu verbessern. Die zentralen Aufgaben betreffen das Verständnis der Voraussetzungen und Effektrichtungen des triebhaften SV-Prozesses, den bewussten Einsatz der Wettbewerbsintensität sowie die Steuerung unterschiedlicher Reaktionen der Leistungsgruppen. Diese Arbeit entwickelt zentrale Gestaltungskriterien, die bei der Konzeption eines RPM-Systems berücksichtigt werden sollten.

Ein wesentlicher Erfolgsfaktor ist ein ausreichend aktivierendes Maß an Ähnlichkeit innerhalb der sozialen Vergleichsprozesse. Wenn dieser Vergleich ausreichend Relevanz besitzt, entsteht eine wirkungsvolle Wettbewerbsumgebung. Innerhalb dieser Steuerung durch Konkurrenz ist die Wahl der Wettbewerbsintensität entscheidend, um kritische Nebeneffekte wie Verzerrung, Rivalität oder Demotivation zu minimieren. Eine höhere Wettbewerbsintensität führt dabei nicht zwangsläufig zu besseren Ergebnissen. Darüber hinaus spielen Persönlichkeitsmerkmale sowie die Art und Dimension der Aufgaben eine zentrale Rolle. Nicht alle Branchen, Berufsgruppen oder Mitarbeitenden reagieren gleichermaßen positiv auf Wettbewerb durch RPM. Besonders komplexe Aufgaben erfordern spezialisierte Gestaltungsansätze, die über einen intuitiven Einsatz hinausgehen. Die differenzierten affektiven Reaktionen der Leistungsgruppen stellen eine weitere Herausforderung dar. Eine gezielte und bedarfsgerechte Gestaltung von RPM kann innerhalb des vielschichtigen Wirkungsprozesses zu einer verbesserten Effektivität führen. Die erfolgreiche Gestaltung von Anreizsystemen auf Basis von RPM verlangt eine sorgfältige Balance zwischen effektiven Vergleichsmaßstäben, der Förderung positiver Wettbewerbsdynamiken und der Berücksichtigung individueller emotionaler Reaktionen.

Für die zukünftige Forschung zu RPM ist eine stärkere Orientierung an einem geordneten Forschungsrahmen notwendig, um verstreute Erkenntnisse zu bündeln und die Mechanismen präziser zu analysieren. Besonders relevant ist die Untersuchung der langfristigen Auswirkungen auf Mitarbeiterbindung und Unternehmenskultur, da diese entscheidend für nachhaltigen Unternehmenserfolg sind. Der Einsatz digitaler Tools und künstlicher Intelligenz eröffnet neue Möglichkeiten, RPM-Systeme dynamisch und individuell anzupassen, um deren Effektivität und Zielgenauigkeit zu steigern. Solche technologiegestützten Ansätze könnten zusätzlich die Vergleichbarkeit und Transparenz verbessern und die Integration von RPM in verschiedene Unternehmenskontexte erleichtern.

Literatur

A. Dye, R. (1992). Relative Performance Evaluation and Project Selection. *Journal of Accounting Research*, *30*(1), 27–52.

A. Hogg, M., & J. Terry, D. (2000). Social Identity and Self-Categorization Processes in Organizational Contexts. *The Academy of Management Review*, *25*(1), 121–140.

A. Long, K. (2021). Internal Amazon Documents Shed Light on How Company Pressures Out 6% of Office Workers. *The Seattle Times Online*. https://www.seattletimes.com/business/amazon/internal-amazon-documents-shed-light-on-how-company-pressures-out-6-of-office-workers/

A. Merchant, K., & T. Otley, D. (2006). A Review of the Literature on Control and Accountability. *Handbooks of Management Accounting Research*, *2*, 785–802.

A. Merchant, K., & Van der Stede, W. A. (2017). *Management Control Systems: Performance Measurement, Evaluation and Incentives* (5. Aufl.). Harlow.

A. Mohnen, K. Pokorny & D. Sliwka. (2008). Transparency, Inequity Aversion, and the Dynamics of Peer Pressure in Teams: Theory and Evidence. *Journal of Labor Economics*, *26*(4), 693–720.

A. Tesser. (1988). Toward a Self-Evaluation Maintenance Model of Social Behavior. In *Advances in Experimental Social Psychology* (S. 181–227, Bd. 21).

A. Tesser, J. Pilkington, C., & D. McIntosh, W. (1989). Self-Evaluation Maintenance and the Mediational Role of Emotion: The Perception of Friends and Strangers. *Journal of Personality and Social Psychology*, *57*(3), 442–456.

A. Tran & R. Zeckhauser. (2012). Rank as an Inherent Incentive: Evidence from a Field Experiment. *Journal of Public Economics*, *96*(9–10), 645–650.

A. Wijayanti, M. Sholihin, E. Nahartyo & Supriyadi. (2024). What Do We Know About the Forced Distribution System: A Systematic Literature Review and Opportunities for Future Research. *Management Review Quarterly*, 1–42.

B. Holmstrom. (1982). Moral Hazard in Teams. *The Bell Journal of Economics*, *13*(2), 324–340.

B. Lount, R., & W. Phillips, K. (2007). Working Harder with the Out-Group: The Impact of Social Category Diversity on Motivation Gains. *Organizational Behavior and Human Decision Processes*, *103*(2), 214–224.

C. Kleine & E. Weißenberger, B. (2014). Leadership Impact on Organizational Commitment: The Mediating Role of Management Control Systems Choice. *Journal of Management Control*, *24*(3), 241–266.

C. Schnieder. (2022). How Relative Performance Information Affects Employee Behavior: A Systematic Review of Empirical Research. *Journal of Accounting Literature*, *44*(1), 72–107.

D. Gill, Z. Kissová, J. Lee & V. Prowse. (2019). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science*, *65*(2), 494–507.

D. Goffin, R., R. Gellatly, I., N. Jackson, D., P. Meyer, J., & V. Paunonen, S. (1996). Criterion Validation of Two Approaches to Performance Appraisal: The Behavioral Observation Scale and the Relative Percentile Method. *Journal of Business and Psychology*, *11*(1), 23–33.

D. Mahlendorf, M., F. Kleinschmit & P. Perego. (2014). Relational Effects of Relative Performance Information: The Role of Professional Identity. *Accounting, Organizations and Society*, *39*(5), 331–347.

D. Otley. (1999). Performance Management: A Framework for Management Control Systems Research. *Management Accounting Research*, *10*(4), 363–382.

D. Tafkov, I. (2013). Private and Public Relative Performance Information under Different Compensation Contracts. *The Accounting Review*, *88*(1), 327–350.

E. Bonner, S. (1999). Judgment and Decision-Making Research in Accounting. *Accounting Horizons*, *13*(4), 385–398.

E. Bonner, S., & B. Sprinkle, G. (2002). The Effects of Monetary Incentives on Effort and Task Performance: Theories, Evidence, and a Framework for Research. *Accounting, Organizations and Society*, *27*(4–5), 303–345.

E. Cardinaels & C. Feichter. (2023). Forced Rankings: Fluch oder Segen? *Controlling und Management Review*, *67*(3), 14–21.

E. Carroll & D. Marginson. (2021). Relative Performance Information and Social Comparisons: Exploring Managers' Cognitive, Emotional and Dysfunctional Behavioral Processes. *Management Accounting Research*, *53*, 1–15.

E. Dechenaux, D. Kovenock & M. Sheremeta, R. (2015). A Survey of Experimental Research on Contests, All-Pay Auctions and Tournaments. *Experimental Economics*, *18*(4), 609–669.

E. Duffy, K., & E. Webber, R. (1974). On "Relative" Rating Systems. *Personnel Psychology*, *27*(2), 307–311.

E. Goodhart, C. A. (1984). Problems of Monetary Management: The UK Experience. In C.A.E. Goodhart (Hrsg.), *Monetary Theory and Practice* (S. 91–121).

E. Strauß & C. Zecher. (2013). Management Control Systems: A Review. *Journal of Management Control*, *23*(4), 233–268.

E. Weick, K. (2015). *Der Prozess des Organisierens* (6. Aufl.).

E. Weißenberger, B. (2007). *IFRS für Controller: Einführung, Anwendung, Fallbeispiele* (1. Aufl.).

E. Weißenberger, B. (2019). Barbara E. Weißenberger Recommends "Management Control Systems: Performance Measurement, Evaluation, and Incentives" by Kenneth A. Merchant and Wim A. Van der Stede. In B. S. Frey & C. A. Schaltegger (Hrsg.), *21st Century Economics. Economic Ideas You Should Read and Remember* (S. 153–155).

F. Hartmann & P. Schreck. (2018). Rankings, Performance, and Sabotage: The Moderating Effects of Target Setting. *European Accounting Review*, *27*(2), 363–382.

F. Ridgway, V. (1956). Dysfunctional Consequences of Performance Measurements. *Administrative Science Quarterly*, *1*(2), 240–247.

F. Thomas, T., & A. Thornock, T. (2015). Me vs. We: The Effect of Incomplete Team Member Feedback on Cooperation of Self-Regarding Individuals. *SSRN Journal*, 1–37.

G. Birnberg, J., J. Luft & D. Shields, M. (2006). Psychology Theory in Management Accounting Research. In *Handbooks of Management Accounting Research* (S. 113–135, Bd. 1).

G. Charness, D. Masclet & C. Villeval, M. (2014). The Dark Side of Competition for Status. *Management Science*, *60*(1), 38–55.

G. Harkins, S., & K. Szymanski. (1989). Social Loafing and Group Evaluation. *Journal of Personality and Social Psychology*, *56*(6), 934–941.

G. Harkins, S., & M. Jackson, J. (1985). The Role of Evaluation in Eliminating Social Loafing. *Personality and Social Psychology Bulletin*, *11*(4), 457–465.

H. Aguinis & K. J. Bradley. (2015). The Secret Sauce for Organizational Success. *Organizational Dynamics*, *44*(3), 161–168.

H. Beach, S. R., & A. Tesser. (1995). Self-Esteem and the Extended Self-Evaluation Maintenance Model. In M. H. Kernis (Hrsg.), *Efficacy, Agency, and Self-Esteem* (S. 145–170).

H. Beth & M. Sasse, C. (2008). Forced Ranking: A Review. *S.A.M. Advanced Management Journal*, *73*(2), 35–39.

H. Eyring & G. Narayanan, V. (2018). Performance Effects of Setting a High Reference Point for Peer-Performance Comparison. *Journal of Accounting Research*, *56*(2), 581–615.

H. Smith, R. (2000). Assimilative and Contrastive Emotional Reactions to Upward and Downward Social Comparisons. In J. M. Suls & L. Wheeler (Hrsg.), *Handbook of Social Comparison: Theory and Research* (S. 173–200).

H. Song, L. Murrell, K., L. Tucker, A., & R. Vinson, D. (2018). Closing the Productivity Gap: Improving Worker Productivity Through Public Relative Performance Feedback and Validation of Best Practices. *Management Science*, *64*(6), 2628–2649.

H. Wruck, K., & Y. Wu. (2022). Relative Versus Absolute Performance Evaluation and CEO Decision-Making. *Journal of Financial and Quantitative Analysis*, *57*(7), 2591–2626.

J. Alsever. (2008). *What Is Forced Ranking?* CBS News. Verfügbar 17. März 2025 unter https://www.cbsnews.com/news/what-is-forced-ranking/

J. Brown, D., L. Ferris, D., D. Heller & M. Keeping, L. (2007). Antecedents and Consequences of the Frequency of Upward and Downward Social Comparisons at Work. *Organizational Behavior and Human Decision Processes*, *102*(1), 59–75.

J. Luft. (2016). Cooperation and Competition Among Employees: Experimental Evidence on the Role of Management Control Systems. *Management Accounting Research*, *31*, 75–85.

J. Nalebuff, B., & E. Stiglitz, J. (1983). Prizes and Incentives: Towards a General Theory of Compensation and Competition. *The Bell Journal of Economics*, *14*(1), 21–43.

J. Suls, R. Martin & L. Wheeler. (2002). Social Comparison: Why, With Whom, and With What Effect? *Current Directions in Psychological Science*, *11*(5), 159–163.

K. Murphy & A. DeNisi. (2023). New Approaches to Dealing with Performance Management: Getting Rid of Performance Appraisals Is Not the Answer. *IIM Ranchi Journal of Management Studies*, *2*(2), 143–158.

K. R. Murphy. (2020). Performance Evaluation Will Not Die, but It Should. *Human Resource Management Journal*, *30*(1), 13–31.

L. Berger, K. Fiolleau & C. MacTavish. (2019). I Know Something You Don't Know: The Effect of Relative Performance Information and Individual Performance Incentives on Knowledge Sharing. *Journal of Management Accounting Research*, *31*(2), 19–35.

L. Festinger. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140.

L. Gerhards & N. Siemer. (2016). The Impact of Private and Public Feedback on Worker Performance-Evidence from the Lab. *Economic Inquiry*, *54*(2), 1188–1201.

L. Hannan, R., P. McPhee, G., H. Newman, A., & D. Tafkov, I. (2013). The Effect of Relative Performance Information on Performance and Effort Allocation in a Multi-Task Environment. *The Accounting Review*, *88*(2), 553–575.

L. Hannan, R., P. McPhee, G., H. Newman, A., D. Tafkov, I., & J. Kachelmeier, S. (2019). The Informativeness of Relative Performance Information and Its Effect on Effort Allocation in a Multitask Environment. *Contemporary Accounting Research*, *36*(3), 1607–1633.

L. Hannan, R., R. Krishnan & H. Newman, A. (2008). The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans. *The Accounting Review*, *83*(4), 893–913.

M. Garcia, S., & A. Tor. (2007). Rankings, Standards, and Competition: Task vs. Scale Comparisons. *Organizational Behavior and Human Decision Processes*, *102*(1), 95–108.

M. Garcia, S., M. Schiff, T., & A. Tor. (2013). The Psychology of Competition: A Social Comparison Perspective. *Perspect Psychol Sci*, *8*(6), 634–650.

N. Eichhorn. (2016). Relative Performance Information and Financial Incentives in Multidimensional Task Settings – A Conceptual and Experimental Analysis of Effects on Performance and Attention Towards Task Dimensions. *Junior Management Science*, *1*, 100–137.

N. Janakiraman, S., A. Lambert, R., & F. Larcker, D. (1992). An Empirical Investigation of the Relative Performance Evaluation Hypothesis. *Journal of Accounting Research*, *30*(1), 53–69.

N. Kluger, A., & A. DeNisi. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, *119*(2), 254–284.

P. Baker, G. (1992). Incentive Contracts and Performance Measurement. *Journal of Political Economy*, *100*(3), 598–614.

P. Buunk, B., I. Gosalvez, A. Nauta, M. Peiro, J., & R. Zurriaga. (2005). Social Comparisons at Work as Related to a Cooperative Social Climate and to Individual Differences in Social Comparison Orientation. *Applied Psychology*, *54*(1), 61–80.

P. Casas-Arce & A. Martínez-Jerez, F. (2009). Relative Performance Compensation, Contests, and Dynamic Incentives. *Management Science*, *55*(8), 1306–1320.
P. Lazear, E., & S. Rosen. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy*, *89*(5), 841–864.
R. Frederickson, J. (1992). Relative Performance Information: The Effects of Common Uncertainty and Contract Type on Agent Effort. *The Accounting Review*, *67*(4), 647–669.
R. J. Akerlof & R. T. Holden. (2012). The Nature of Tournaments. *Economic Theory*, *51*(2), 289–313.
R. Leary, M., & M. Kowalski, R. (1990). Impression Management: A Literature Review and Two-Component Model. *Psychological Bulletin*, *107*(1), 34–47.
S. Adler, M. Campion, A. Colquitt, A. Grubb, K. Murphy, R. Ollander-Krane & E. D. Pulakos. (2016). Getting Rid of Performance Ratings: Genius or Folly? A Debate. *Industrial and Organizational Psychology*, *9*(2), 219–252.
S. Blader, C. Gartenberg & A. Prat. (2019). The Contingent Effect of Management Practices. *The Review of Economic Studies*, *87*(2), 721–749.
S. Goebel & E. Weißenberger, B. (2017). Effects of Management Control Mechanisms: Towards a More Comprehensive Analysis. *Journal of Business Economics*, *87*(2), 185–219.
S. Kramer, S. Maas, V., & V. Rinsum, M. (2016). Relative Performance Information, Rank Ordering and Employee Performance: A Research Note. *Management Accounting Research*, *33*, 16–24.
S. Murthy, U. (2010). The Effect of Relative Performance Information Under Different Incentive Schemes on Performance in a Production Task. *SSRN Journal*, 1–35.
S. Murthy, U., & A. Schafer, B. (2011). The Effects of Relative Performance Information and Framed Information Systems Feedback on Performance in a Production Task. *Journal of Information Systems*, *25*(1), 159–184.
S. Stewart, M. Gruys & M. Storm. (2010). Forced Distribution Performance Evaluation Systems: Advantages, Disadvantages, and Keys to Implementation. *Journal of Management and Organization*, *16*, 168–179.
T. Baldwin, T., D. Blume, B., & S. Rubin, R. (2013). Who Is Attracted to an Organisation Using a Forced Distribution Performance Management System? *Human Resource Management Journal*, *23*(4), 360–378.
T. Ball, R., J. Bonham & T. Hemmer. (2020). Does It Pay to Be Like Mike? Aspirational Peer Firms and Relative Performance Evaluation. *Review of Accounting Studies*, *25*(4), 1507–1541.
W. Chan, E. (2018). Promotion, Relative Performance Information, and the Peter Principle. *The Accounting Review*, *93*(3), 83–103.
W. Thorngate. (1976). "In General" vs. "It Depends": Some Comments of the Gergen-Schlenker Debate. *Personality and Social Psychology Bulletin*, *2*(4), 404–410.
W. Wang, L. (2017). Recognizing the Best: The Productive and Counterproductive Effects of Relative Performance Recognition. *Contemporary Accounting Research*, *34*(2), 966–990.
Y. Chen, M. Harper, F., J. Konstan & X. Li, S. (2010). Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review*, *100*(4), 1358–1398.

Abkürzungsverzeichnis

APM Absolute Performance-Messung

AT Agency-Theorie

FR Forced-Ranking

MC Management-Control

MCS Management-Control-System

MH Moral Hazard

OoC Object-of-Control

PM Performance-Messung

RPI Relative Performance-Informationen

RPM Relative Performance-Messung

SV Sozialer Vergleich

TSV Theorie des sozialen Vergleichs

Symbolverzeichnis

A Bereitstellungsparameter

B Personenbezogene Variablen

$c(e_i)$ Kostenfunktion des Agenten i für die erbrachte Anstrengung

C Aufgabenbezogene Variablen

D Umgebungsbezogene Variablen

e_i Anstrengung des Agenten i

ε_c Systemische Unsicherheitsvariable

ε_i Idiosynkratische Unsicherheitsvariable

$f_e(A, B, C, D)$ Anstrengungsvermittelte Performancefunktion

k Kostenempfindlichkeit der Anstrengung

L Fixe Grundvergütung

μ_1 Mittelwert der Normalverteilung

μ_2 Mittelwert der Potenzgesetzverteilung

$\mathbb{P}(x_i > x_j)$ Gewinnwahrscheinlichkeit für den Turnierpreis

U_i Nutzenfunktion des Agenten

W Variabler Turnierpreis

X Gesamtperformance

x_i; x_j Performance eines individuellen Agenten