IT UNIVERSITY OF COPENHAGEN

Designing Artificial Intelligence Systems for B2B Aftersales Decision Support

A PhD Thesis

by Tiemo Thiess, IT University of Copenhagen

Supervisors:

Principal Supervisor: Dr. Oliver Müller, IT University of Copenhagen

Co-Supervisor: Dr. Raffaele Fabio Ciriello, The University of Sidney

Company Supervisor: Oliver Mühlich, MAN Energy Solutions

Company Co-Supervisor: Szymon Furtak, MAN Energy Solutions

Committee:

Dr. Oliver Krancher, IT University of Copenhagen

Dr. Iris Junglas, College of Charleston

Dr. Guido Schryen, Paderborn University

To Imke, my Family, and Friends.

Acknowledgment

I want to thank Dr. Oliver Müller for always supporting me, giving me all the creative space I wanted, and all the guidance I needed, for inspiring me with his engaging, informative, and yet entertaining teaching, and with his precise and refreshing writing style. Most of all, I want to thank him for being the best supervisor that I could have wished for, despite opposing preferences for football clubs. Also, I want to thank Dr. Raffaele Ciriello for supporting me as a co-supervisor.

I want to thank MAN Energy Solutions for co-financing this PhD project, and for allowing me to study their organization, but also for giving me the chance to design and shape novel AI systems for B2B aftersales support in a natural field setting. Especially, I want to thank Oliver Mühlich for being a reliable, supportive, and great company supervisor, who, despite the often more pragmatic objectives of private organizations, supported my scientific work fully. Also, I want to thank Harald Capek for making this unique industry collaboration possible and for supporting me in all bureaucratic tasks, but also in helping me to gain organizational momentum, anchorage, and management support. I want to thank Gert Steen Sørensen, the data wizard of MAN, for helping me with all data-related questions and issues, and for sharing some of his mighty SQL spells. I want to thank Maria Colberg Weiss for helping me create data-driven leads in MAN's customer relationship management system and for being fun and great to work with. Moreover, I want to thank Lorenzo Tonelli for his support in developing and implementing an explainable AI system for aftersales quote follow-up, and for demonstrating what a real business-technology boundary spanner is. Furthermore, I want to thank Dr. Stefan Schneider for the great collaboration in building a causal inference system for spare parts pricing effect estimation. And I want to thank Szymon Furtak for being my co-supervisor at MAN Energy Solutions.

I also want to thank the Innovation Fund Denmark for co-financing this collaborative PhD project and for always being supportive when I had questions and with that assuring smooth and steady project progress.

Moreover, I want to thank Arisa Shollo for making my research stay at the Digitalization Department at Copenhagen Business School possible, and for being, just like Dr. Konstantin Hopf, a great colleague, and co-author to work with.

I want to thank my friends, especially, Christoffer, for his help with the Danish abstract, and Christoph, Per, and Hissu, for making my PhD journey a great experience. Finally, I want to thank my family, for always believing in me, and Imke, for being the kind, positive, and loving person she is.

English Abstract

Original equipment manufacturers have started to transform and servitize their business models and processes. This transformation entails shifting their focus from the product development towards the product usage phase. In this phase, they can monetize aftersales services such as maintenance, repair, overhaul, and spare parts delivery. However, low-cost competitors threaten the potential gains from this transformation. This threat requires original equipment manufacturers to become more customer-centric and exploit their internal resources better. Artificial intelligence (AI) has the potential to enable such customer-centric B2B service strategies. However, especially in B2B contexts, professionals and researchers lack guidance about how to design and implement effective AI systems.

In reaction to this, I conducted three action design research (ADR) studies at MAN Energy Solutions that follow the dual-mission of information systems research, to create utility for practitioners while extending the scientific body of knowledge. These ADR studies resulted in three implementations of novel AI systems. The systems represent the state-of-the-art in AI value creation while addressing the specific challenges of B2B aftersales contexts. In addition to this, I developed a set of design principles that explicitly guide practitioners and researchers on how to design and implement AI systems for B2B aftersales decision support.

These AI systems, in turn, enable B2B firms in general and original equipment manufacturers, in particular, to adopt technology-driven and customer-centric service strategies and thereby to create competitive advantages by providing personalized and high-quality service that the low-cost competition cannot provide.

Dansk Abstrakt

Producenter af originalt udstyr er begyndt at transformere deres forretningsmodeller og processer. Denne transformation medfører, at deres fokus flyttes fra produktudviklingen mod produktforbrugsfasen. I denne fase kan de tjene penge på eftersalgstjenester som vedligeholdelse, reparation, eftersyn og levering af reservedele. Men lav-pris konkurrenter truer de potentielle gevinster ved denne transformation. Denne trussel kræver, at producenter af originalt udstyr bliver mere kundecentriske og udnytter deres interne ressourcer bedre. Kunstig intelligens (AI) har potentialet til at muliggøre sådanne kundecentriske B2B-servicestrategier. Især i B2B-sammenhænge mangler fagfolk og forskere vejledning til, hvordan man designer og implementerer effektive AI-systemer.

Som reaktion på dette gennemførte jeg tre action-design-undersøgelser (ADR) på MAN Energy Solutions, der følger den dobbelte mission af informationssystemsforskning, for at skabe brugbarhed for udøvende fagfolk, mens man udvider den videnskabelige vidensbase. Disse ADR-undersøgelser resulterede i tre implementeringer af nye AI-systemer. Systemerne repræsenterer det nyeste inden for AI-værdiskabelse, og imødekommer de specifikke udfordringer i forbindelse med B2B-eftersalg. Derudover har jeg udviklet et sæt designprincipper, der eksplicit vejleder de udøvende fagfolk og forskere i, hvordan man designer og implementerer AI-systemer som vil understøtte B2B eftersalgsbeslutningsprocessen.

Disse AI-systemer gør det både muligt for B2B-virksomheder i almindelighed, og i særdeleshed for producenter af originalt udstyr, at adoptere teknologidrevne og kundecentriske servicestrategier. Derved kan de skabe konkurrencefordele, ved at levere personlig service af høj kvalitet, som lav-pris konkurrenter ikke kan levere.

Contents

Part A

1 Introduction

Original equipment manufacturers (OEMs) are undergoing a transformational process of servitizing their business processes and models (Baines et al. 2017; Lightfoot et al. 2013; Luoto et al. 2017). For OEMs, this means shifting the focus from product development and selling physical products only to the more customer-focused product usage phase (aftersales) in which one can monetize, e.g., spare parts, repair, and maintenance services (Sundin 2009). Baines and Lightfoot (2012) differentiated three levels of aftersales services and servitization: "base" (e.g., product installation and spare parts provisioning), "intermediate" (maintenance, repair, and overhaul), "advanced" (service contracts, product output focused-contracts). Research suggests that many OEMs have successfully established structures to capitalize on the base level of aftersales services and especially on provisioning (selling) spare parts to their main-product customers, but that advanced services are rarely offered (Adrodegari et al. 2014).

Nevertheless, already the first stage of servitization is potentially highly profitable. However, competition from numerous and often small third party companies selling non-original spare parts at low prices threatens such profits significantly (Gallagher et al. 2005). However, compared to this low-price competition, OEMs have many resources that they can turn into competitive advantages. They usually have stronger customer relationships, better supply chains, stronger engineering capabilities, more knowledgeable technical support, and, generally, higher product quality, and quality assurance (Gallagher et al. 2005). Nevertheless, OEMs struggle to exploit these resources fully (Cohen et al. 2006).

Research suggests that instead of competing on low-prices, OEMs should exploit their resources to increase their service quality (Cohen et al. 2006; Gallagher et al. 2005). This view is equivalent to perceiving customers as intangible assets that one should invest in so that they not only generate cash-flow in the present (transaction focus) but also in the future (relationship focus; Blattberg et al. 2001; Gupta and Lehmann 2005). However, despite the resource advantage that OEMs have, they cannot drastically increase service quality for all customers. Therefore, "[a] high level of service for one customer may, therefore, necessitate a lower level of service for another"(Cohen et al. 2006). Instead, research suggests that firms should invest only in those

customers that have high future value for the firm (Ascarza et al. 2017; Fader 2012). However, to identify those customers accurately, "the firm must have access to rich customer-level data from both internal and external sources, along with the capabilities to analyze these data" (Ascarza et al. 2017, p. 2).

Huang and Rust (2017), on the other hand, argue that increased availability of data along with new digital technologies and particularly AI, not only enables to identify high potential customers but to approach all customers with nuanced "technology-driven service strategies". In particular, they suggest that firms that have customers with heterogeneous demand patterns, such as OEMs in the aftersales market (Bartezzaghi et al. 1999; Bartezzaghi and Kalchschmidt 2011), should exploit AI and big data resources to design personalized instead of standardized services. Moreover, such firms should approach high potential customers with a "dynamic personalization" strategy while they should approach low potential customers with a "static personalization" strategy. Examples of static personalization would be to use customer segmentation to send targeted newsletters or advertising or to adjust the design of an e-commerce platform based on a customer's segment. A static personalization strategy makes only sense if many low-potential customers can be approached efficiently, e.g., via an e-commerce platform.

Conversely, a dynamic personalization strategy could involve frontline worker:

> In contrast to static personalization, dynamic personalization adapts to a specific customer's preferences based on his/her active input and based on observing his/her behavior over time, rather than relying mostly on cross-sectional like-minded customer data inference. The longer and deeper the relationship, the better the personalization can be adapted over time. (Huang and Rust 2017, p. 12)

AI is especially suitable for such dynamic personalization strategies as "[c]ompared to the more static data- and text-mining techniques that focus on discovering patterns from a large scale of cross-sectional data, AI is more adaptive and focuses on learning from an individual customer's past behavior" (Huang and Rust 2017).

Nevertheless, while some business to consumer (B2C) organizations have already established such customer-centric and data-driven approaches, they are less established in business to business (B2B) and specifically OEM aftersales contexts (Martínez-López and Casillas 2013; Mora Cortez and Johnston 2017; Stormi et al. 2018). Martínez-López and Casillas (2013) call for more scientific contributions on applications of AI in industrial marketing contexts and Mora Cortez and Johnston (2017) criticize the B2C focus in scientific outlets and argue that there is a gap between the real-work challenges that B2B practitioners face and the guidance that B2B theory can provide. They propose data analytics as one of the key areas of research that B2B marketing should focus on to "resolve real problems that B2B marketers will face during the next three to five years" (p. 6). In particular, they call for "[e]stablishing procedures for data gathering and analysis to improve decision-making" (p. 9).

Based on the discussion above, I motivate the first research question of this PhD project with the high potential of AI systems for enabling customer-centric aftersales service processes and personalized technology-driven service strategies. Lacking research that focuses on how to design AI systems and procedures for data-driven decision support in B2B marketing motivates the research question further:

RQ1: How to design AI systems that support B2B aftersales processes and strategies?

To answer the research question, I defined three research goals:

RG1: To design AI systems that are informed by state-of-the-art knowledge about AI value creation

RG2: To design AI systems addressing the particular challenges of B2B aftersales contexts

RG3: To develop design principles that guide designers in constructing artifacts of the class of AI systems for B2B aftersales decision support in particular, and, where applicable, the broader class of AI decision support systems

This PhD project approaches these research goals and the research question by conducting an action design research (ADR) program at MAN Energy Solution, one of the leading OEMs for large two- and four-stroke diesel engines. To fulfill RG1, the PhD project started with an inten-

sive study of the existing literature on AI value generation mechanisms. The existing research shows strong evidence for the superiority of AI decision making over human decision making (Ægisdóttir et al. 2006; Dawes 1979; Dawes et al. 1989; Grove et al. 2000; Grove and Meehl 1996; Meehl 1954). Also, studies could demonstrate the effects of using AI systems on an organizational level (Brynjolfsson et al. 2011; Müller et al. 2018). These studies focus on the effects of the technology itself and do not say much about the social-technical complexities that arise when AI artifacts are deployed in organizations (see Sharma et al. 2014). Sharma et al. (2014) argue that AI and analytics value creation involves circumventing challenges not only in converting data to insights, but in converting such insights into decisions, and such decisions into actions and value. They call for research that investigates the role that organizational structures and human decision making processes play in the design and implementation of AI and analytics systems and how such systems shape their organizational environment. During the initial literature review on AI value creation mechanisms, it became apparent that the research gap that Sharma et al. (2014) identified was still not addressed in a way that is suitable to inform the design of effective AI solution artifacts. Therefore, the PhD project follows a second research question, which helps to fill this research gap but eventually guides fulfilling RG1 in the overall ADR program.

RQ2: How can organizations create and sustain value through applications of AI?

The PhD project follows RG4 to answer RQ2.

RG4: To create behavioral knowledge about how real-world designers approach the holistic data-to-insights-to-decisions-to-actions-to-value path with organizational applications of AI

The ADR program consisted of three separate ADR studies, in which we designed and implemented three AI systems that help to solve relevant field problems along the B2B aftersales service funnel. In the first ADR study, we designed and implemented an AI system for data-driven lead generation that enables pro-active customer relationship processes and improved conversion rates from leads to requests for aftersales quotations. In the second ADR study, we designed and implemented an AI system that predicts the win-propensity of aftersales quotations,

and thereby, supports sales professionals in their quotation follow-up processes. In the third ADR study, we designed and implemented an AI system that supports pricing analysts in estimating the impact of value-based pricing strategies on aftersales orders. Overall, ADR Study 2 and ADR Study 3 support the conversion process of aftersales quotations into aftersales orders.

A behavioral study that we conducted in parallel to the ADR studies informs all three ADR projects of key challenges, enablers, and mechanisms in AI value creation. Figure 1 shows the overall timeline of the different studies that comprise this PhD project. We started ADR study 1 already before starting the PhD project. My unpublished master thesis reports on the first iterations of the designed data-driven lead generation artifact. The two papers about this study, which are part of this PhD thesis, report on further iterations of the artifact, and present a much more developed conceptualization of design principles. Figure 1 also shows that we conducted the AI value creation study in parallel to the three ADR studies. While it mostly informed the conceptualization of design principles for ADR Study 1, for the two other ADR studies, it guided the decision of the mayor theoretical bases that we inscribed into the artifacts, namely interpretable AI, and causality. Moreover, these behavioral results structure the discussion of the designed AI systems and their underlying design principles.

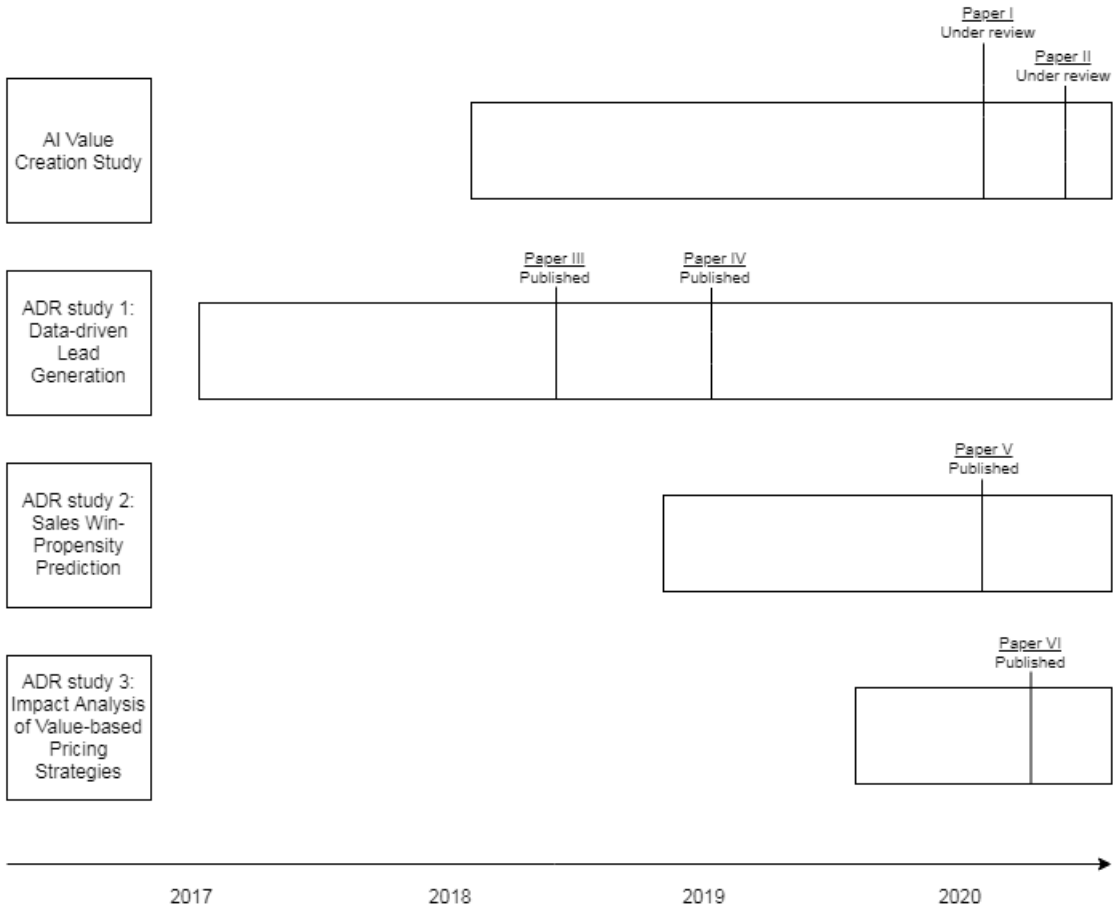

Figure 1. Timeline studies

The thesis is structured as follows. In Section 2, it discusses the theoretical background in further detail. Section 3 describes the B2B aftersales context at MAN Energy Solutions. Section 4 further describes the ADR methodology that the overall research program followed. Section 5 summarizes the main research results of the different sub-studies that comprise this cumulative PhD thesis. Finally, Section 6 discusses the overall contributions of this PhD project, the implications for research, practice, some limitations, and concludes Part A. Part B presents the papers related to the different studies in full detail.

Table 1 lists the publication, outlet, ranking, status, study, and the research goals that the study aims to achieve. In the ranking column, I show the bibliometric research indicator (BFI), which is a fundamental part of the Danish university system. Based on the BFI, one makes decisions about funding allocations to universities and researchers. The BFI gives points between 1 and 3

for original peer-reviewed publications. Here, 1 stands for ordinary publications, 2 for distinguished, and 3 for excellent and most prestigious. Moreover, I show the JOURQUAL3 (JQ3) ranking for information systems research provided by the Verband Deutscher Hochschullehrer. Here A stands for a leading scientific outlet, B for an important scientific outlet, C for a respected scientific outlet, and D for a scientific outlet. Moreover, the column shows the ranking of the Wissenschaftliche Kommission für Wirtschaftsinformatik (WKWI) that has the same interpretation as the JQ3, and, when accessible, the acceptance rate (AR) of the outlet.

Table 1. Publication Overview

Publication	Outlet	Ranking	Status	Study	Goal(s)
Müller, Thiess, Hopf, and Shollo, "Challenges and Enablers along the AI Value Creation Process"	*California Management Review (CMR)*	BFI: 2, JQ3: B	Under review	AI Value Creation	1 and 4
Shollo, Hopf, Thiess, and Müller, "Shifting AI Value Creation Mechanisms: An Explorative Study"	*The Journal of Strategic Information Systems (JSIS)*	BFI: 2, JQ3: A, WKWI: A	Under review	AI Value Creation	1 and 4
Thiess and Müller 2018, "Towards Design Principles for Data-driven Decision Making – An Action Design Research Project in the Maritime Industry"	*Proceedings of the European Conference on Information Systems (ECIS)*	BFI: 1, JQ3: B, WKWI: A, AR: 30%	Published	ADR Study 1: Data-driven lead generation	1-3
Thiess and Müller 2020a, "Setting Sail for Data-driven Decision Making – An Action Design Research Case from the Maritime Industry"	*Design Science Research. Cases (Springer Series Progress in IS)*	BFI: 2	Published	ADR Study 1: Data-driven lead generation	1-3
Thiess et al. 2020, "Design Principles for Explainable Sales Win-Propensity Prediction Systems"	*Proceedings of the Internationale Tagung Wirtschaftsinformatik (WI)*	JQ3: C, WKWI: A	Published	ADR Study 2: Sales Win-Propensity Scoring	1-3
Thiess and Müller 2020b, "Designing Causal Inference Systems for Value-based Spare Parts Pricing – An ADR Study at MAN Energy Solutions"	*Lecture Notes in Business Information Processing (LNBIP)*	BFI: 1, VHB: C, AR: 29%	Published	ADR Study 3: Causal Impact analysis of value-based pricing strategies	1-3

2 Theoretical Background

This chapter discusses the theoretical background of this thesis. It gives a background to the AI Value Creation Study but mostly discusses the theory that informed the design of AI systems 1, 2, and 3 that are the focus of ADR studies 1, 2, and 3. In particular, 2.1 discusses essential concepts and terms such as artificial intelligence, machine learning, and data scientist that repeatedly appear in this thesis. As the AI systems that this thesis presents support human decision making, Section 2.2 discusses human decision making processes and capabilities and how AI can augment them. Section 2.3 discusses the theory that directly informed the problem formulation and solution design of AI System 2 and AI System 3.

2.1 Artificial Intelligence, Machine Learning, and Related Concepts

The recent surge in artificial intelligence (AI) related breakthroughs like self-driving cars (Brynjolfsson and Mitchell 2017) or advanced disease diagnostics (McKinney et al. 2020) has put it on the agenda of scientific debate. While many definitions of AI exist, this PhD thesis follows a well-accepted definition (more than 35.000 citations) by Russel and Norvig (2009) who perceive AI as information technology (IT) based artifacts that either think or act like humans or think or act like rational agents. Here they base their definition of *thinking like humans* on Bellman (1978), who argues that it involves "activities that we associate with human thinking, activities such as decision-making, problem-solving, learning." Moreover, *acting like humans* was defined as "to make computers do things at which, at the moment, people are better" (Rich and Knight 1990). While *thinking like rational agents* was defined as "computations that make it possible to perceive, reason, and act" (Winston 1992). Moreover, *acting like rational agents* refers to "intelligent behavior in artifacts" (Nilsson 1998).

The first successful AI applications were so-called expert systems, which mimic human experts (Russel and Norvig 2009). They are logic-based and consist of numerous what-if rules that, when applied on data, can give new answers. However, this PhD focusses on the human-like thinking capabilities of AI technologies and, in particular, on machine learning. One can distinguish different machine learning types. In supervised learning, algorithms try to find a function

$h(x) = y$ based on input vectors that contain a set of features (x) and historical observations of the target (y; Russel and Norvig 2009). Here, supervised learning of a discrete target variable is a classification task, while supervised learning of a continuous target variable is a regression task. Figure 2 illustrates the supervised learning process in ML-based systems and compares this to how non-learning rule-based systems generate outputs in the form of inferences or recommendations based on knowledge and data.

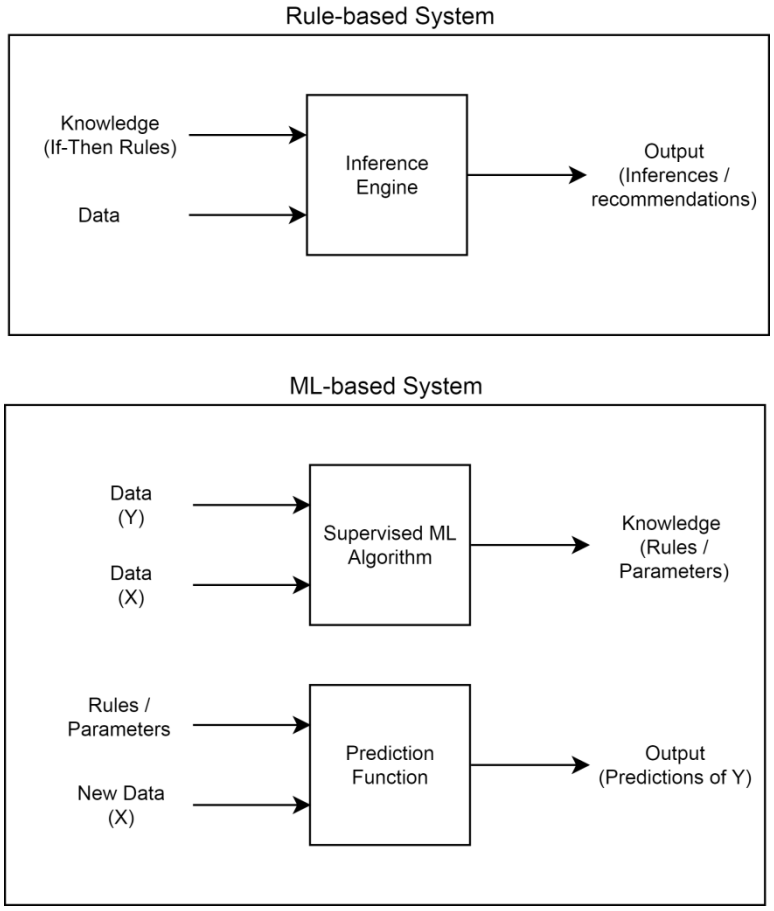

Figure 2. Difference between rule-based and ML-based systems

Other concepts that are related to this PhD thesis are data-driven decision making (DDD), big data, big data analytics, data science, and statistics. Data-driven decision making (DDD), emphasizes the utilization of data and statistical algorithms rather than solely relying on human decision making (Brynjolfsson et al. 2011). According to Provost and Fawcett (2013), data science is based on engineering and processing of data and big data and enables company-wide

DDD. Here, they define data science as "principles, processes, and techniques for understanding phenomena via the (automated) analysis of data" (p. 53). They define big data from a technological viewpoint as "datasets that are too large for traditional data-processing systems and that therefore require new technologies" (p. 54). This PhD thesis follows the above-outlined definitions while perceiving data science as a broader set of activities that are necessary to turn data into value. Consequently, this PhD thesis perceives data scientists as the key organizational actors that design and implement machine learning algorithms to create AI systems and organizational value. Moreover, following Leek and Peng (2015), I differentiate the term statistics (or data analysis) into descriptive statistics (summarizing data without interpretation), inferential statistics (quantifying the relationship between independent and dependent variables on a population level), predictive statistics (using the quantified relationships and new data to predict an unobserved value of the dependent variable for a single unit), and causal statistics (causal inference): "A causal data analysis seeks to find out what happens to one measurement on average if you make another measurement change. Such an analysis identifies both the magnitude and direction of relationships between variables on average" (p. 1315).

In this PhD, I perceive descriptive statistics not as AI, while I perceive inferential and causal statistics as forms of machine learning, and, thus, AI that follow explanatory goals and predictive statistics as a form of machine learning that follows predictive goals (Shmueli 2010). Moreover, in all three cases, one applies supervised learning algorithms such as linear or logistic regression that learn rules and find functions $h(x) = y$ from data in the above-described and in Figure 2 illustrated way.

2.2 Human and AI-augmented Decision Making

Most AI systems support and augment human decision making rather than entirely replacing it. Because of this, the following section discusses, first, research on human decision making under risk (2.2.1) and uncertainty (2.2.2), and, second, how AI augments human decision making (2.2.3).

2.2.1 Human Decision Making under Risk: Objective Rationality

Decision theory suggests that rational decision-makers, when confronted with a decision problem, should choose the option that maximizes a subjective expected utility function. Often people express such utility monetarily (Edwards 1954; Leonard and others 1954; von Neumann and Morgenstern 1944). Nevertheless, in organizational contexts, one can doubt how strictly people follow the normative theories of rational decision theory. Inspired by the von-Neumann-Morgenstern theorem (von Neumann and Morgenstern 1944), Simon (1997) developed an objectively rational decision making process that consists of three steps:

1.) listing of all the alternative strategies
2.) the determination of all the consequences that follow upon each of these strategies
3.) the comparative evaluation of these sets of consequences

However, according to Simon (1997), an individual's objective rationality is limited:

> It is impossible for the behavior of a single, isolated individual to reach any high degree of rationality. The number of alternatives he must explore is so great, the information he would need to evaluate them so vast that even an approximation to objective rationality is hard to conceive. (p. 92)

Based on such realizations, newer views on decision making processes (Bazerman and Moore 2008; Goodwin and Wright 2014) do not include definite terms such as "all" but are more implying that people, for instance, generate "some" alternatives :

1.) problem definition
2.) criteria identification and weighting
3.) alternative generation
4.) alternative rating for each criterion
5.) computation of optimal decision
6.) decision implementation

2.2.2 Human Decision Making under Uncertainty: Satisficing and Bounded Rationality

Several studies have shown the superiority of ML and related approaches over human decision making in numerical prediction tasks (Ægisdóttir et al. 2006; Dawes 1979; Dawes et al. 1989; Grove et al. 2000; Grove and Meehl 1996; Meehl 1954). A possible explanation for this are the "psychological limits of the organism (particularly with respect to computational and predictive ability)" (Simon 1955, p. 101). Simon (1997) perceives individuals' rationality as limited or bounded by their value system, knowledge, skills, and the information that is accessible at the point of decision making. To cope with such limitations, Simon (1979) argues that humans "satisfice either by finding optimum solutions for a simplified world or by finding satisfactory solutions for a more realistic world" (p. 498). He describes the former as "[isolating] from the rest of the world a closed system containing only a limited number of variables and a limited range of consequences […]" (Simon 1997, p. 95), and the latter as an individual's behavior of choosing the first alternative whose valuation satisfies a priorly-set value-threshold, the value-threshold, however, can be lowered if no alternative satisfices it (Simon 1955). He further clarifies, "[o]nly those factors that are most closely […] connected in cause and time can be taken into consideration" (Simon 1997, p. 95).

Based on Simon's seminal work, Kahneman and Tversky (1974) started an intensive study of judgment and decision making under uncertainty with an explicit focus on heuristics and biases. Kahneman (2003) argues that heuristics are mental shortcuts mostly governed by an unconscious, associative, automatic, and effortless thinking (System 1) that stands in contrast with a conscious, rule-based, sequential, and effortful thinking system (System 2; see also Stanovich and West 2003). It is in System 2, where most deliberate reasoning happens. In particular, Kahneman and Frederick (2002) argue that heuristics are explained by the process of attribute substitution, according to their theory, when faced with a cognitively demanding decision problem, humans, substitute it unconsciously with a simpler one. Heuristics, in turn, can be seen as biases since they deviate from the normative models of objectively rational decision making as described above (Tversky and Kahneman 1974).

However, Gigerenzer (2008) argues that in general, heuristics are efficient mechanisms considering the time and resource constraints that decision-makers are faced with in real-life situa-

tions, while Kahneman and Tversky (1974) see biases as deviations from objectively rational decision making, mostly due to their intuitive System 1 characteristic. Gigerenzer's viewpoint is more aligned with Simon's view on bounded rationality (Simon 1955) as he argues for ecological rationality (2008) by stating that the violation of objective rationality was empirically proven to be, in fact, rational in particular problem environments. Gigerenzer (2008) views heuristics as decision strategies that one can use deliberately, thus, governed by Kahneman's (2003) System 2 thinking. In summary, one can argue that heuristics, when used via the intuitive System 1 thinking, are prone to cause decision errors and cognitive biases in light of most views on rationality. When used via the deliberate System 2 thinking, however, they can be a rational decision strategy considering the limitations of both the decision-maker, and the complexity induced by the decision environment.

For Simon, the implications of his theory of bounded rationality (1955; 1997) are that one can improve organizational decision making both by (re-)designing the external environment (e.g., structural complexity, time pressure, group behavior) and the internal environment (e.g., skills, habits). According to Simon, designing is a deliberate problem-solving process: "Everyone designs who devises courses of action aimed at changing existing situations into preferred ones" (Simon 2019).

2.2.3 AI-augmented Decision Making

In deliberate human decision making, several steps require numerical judgments; for instance, when estimating the utility of decision alternatives. Such numerical judgments are predictions since the human anticipates—based on knowledge and information—how likely a future consequence materializes (Kahneman and Frederick 2002).

In the case of augmented decision making, humans can get AI support in a much more structured and reliable way, given that behavioral research has shown that humans are particularly bad at numerical predictions (Kahneman and Frederick 2002). They often replace, for example, the difficult numerical prediction problem with a simpler similarity matching task (heuristics). AI systems, on the other hand, have their core strengths in the accurate prediction of numerical values and, in particular, the estimation of probabilities (see section 2.1.). Therefore, it is no

wonder that several empirical studies report the superiority of AI in prediction problems (e.g., Grove et al., 2000; Grove and Meehl, 1996). Nevertheless, AI systems are still dependent on their human choice architects in the form of data scientists, developers, user interface (UI) designers, etc., that make many subjective decisions, e.g., about data selection, processing, modeling, algorithm selection, UI design, IT architecture and system execution. This PhD thesis perceives such choice architects as the main target group.

2.3 AI Value Creation Challenges: Causes, Consequences, and Enablers

Among the key findings of the AI value creation study (see Section 5.1.) are seven challenges for AI value creation along with a set of enablers. The three AI artifacts developed during this PhD ADR program address four of those. In the following, I discuss the theories that either inspired the initial solution design or justified a solution design for the challenges ex-post. Finding solutions for a problem requires understanding the causes of the problem and its consequences. As the focus of this PhD was on designing novel AI artifacts, I did not conduct explanatory studies for understanding the causes behind each challenge. However, I followed a pragmatic approach and explained the causes in terms of abduction to the best explanation. Therefore, the search for theory that can explain the theoretical and later also practical challenges that this PhD project faced, is not exhaustive but limited by my own knowledge and that of my supervisor and colleagues that I consulted.

2.3.1 The AI Interpretability Challenge and Decision Acceptance

I approached to understand the cause of the Interpretability challenge as the inability to compare one's mental model with an intransparent and black-boxed AI model.

Du et al. (2020, p. 1) argue that “the lack of transparency behind” the “behaviors” of black-boxed AI systems, e.g., based on neural networks, is one of their most severe limitations as it

> leaves users with little understanding of how particular decisions are made by these models. Consider, for instance, [if] an advanced self-driving car equipped with various machine learning algorithms doesn’t break or decelerate when confronting a stopped

> firetruck. This unexpected behavior may frustrate and confuse users, making them wonder why. Even worse, the wrong decisions could cause severe [unintended] consequences if the car is driving at highway speeds and might finally crash the firetruck. The concerns about the black-box nature of complex models have hampered their further applications in our society, especially in those critical decision-making domains.

Research on using AI for decision making has mainly focused on converting data into high-quality insights and decisions. However, to create value from implementing them successfully, high-quality decisions are not enough (Sharma et al. 2014). Based on Maier (1963), Hollander et al. (1973) argued that effective decisions are a function of its quality, its acceptance by the stakeholders that shall implement it, and its complexity (in terms of the time one has to invest in making the decision). Moreover, they argued that the extent to which stakeholders participate in and influence decision making processes increases the acceptance of the eventual decision.

Another important factor that impacts the acceptance of an AI system is the interpretability of its underlying decision making processes because "[if] human decision makers do not know the rationale behind the suggested recommendation, they are typically skeptical of the output produced and are therefore reluctant to use such systems" (Umanath and Vessey 1994, p. 796). Kayande et al. (2009) reasoned "that a lack of user understanding of the logic underlying [decision support system] output leads to poor perceptions of the value such model-based [decision support systems] offer, leading to user resistance and impeded system use" (p. 528). Lilien et al. (2004) emphasized the "perception-reality gap" between objectively making high-quality decisions when using decision support systems and subjectively evaluating such decisions as low-quality. They stated that their "results suggest that what managers get from a [decision support system] may be substantially better than what they see" (p. 216). Based on this, they proposed that one should design decision support systems so that users can align their subjective perception of the systems' quality with its real quality. In particular, they proposed "to design in features that encourage interaction with the [decision support system], offer explanations for recommendations, generate visual outputs, and provide structured cognitive and outcome feedback" (p. 233).

Kayande et al. (2009) introduced the 3-gap framework (see Figure 3). It suggests, firstly, that a gap between the decision maker's mental model and the AI model affects AI system acceptance, secondly, that a gap between the AI model and the true model (reality) affects the AI system performance, and, thirdly, that a gap between the decision maker's mental model and the true model (reality) affects the decision maker's performance. Here, a lack of a decision maker's understanding of the logic behind the recommendations that the AI system produces enlarges the gap. This creates a conflict between the AI "model's recommended course of action and that implied by the user's mental model […] resulting in decision uncertainty" (Kayande et al., 2009, p. 529, referring to Einhorn and Hogarth 1985). Moreover, based on a theory of preferences with risk adjustments (Keeney et al. 1993), Kayande et al. (2009) proposed that in reaction to such decision uncertainty, especially risk-averse decision-makers would subjectively evaluate the quality of the AI systems as lower than it objectively is. Based on this, they suggested that the inability of AI system designs to close Gap 1 between the decision maker's mental model and the AI model is a potential cause of the above-described perception-reality gap in AI system evaluation, and AI system acceptance in general.

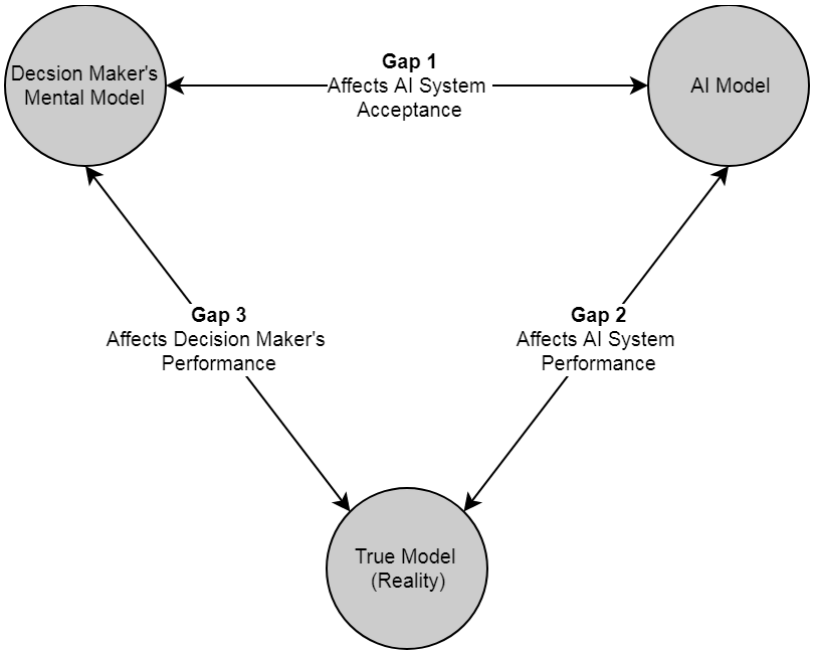

Figure 3. 3-Gap Framework adapted from Kayande et al. (2009)[1]

In their argumentation, Kayande et al. (2009) assumed that Gap 2 is always smaller than Gap 3, and the AI system, therefore, always superior to the decision maker's mental model in terms of decision quality. Martens and Provost (2014) criticized this based on the argument that it may be unrealistic to assume that AI systems are always superior to human decision-makers in terms of decision quality, due to the introduction of biases during the model building process, or overfitting a model on training data. In reaction to this, they extend the 3-gap framework, e.g., by adding with developers, managers, and clients different roles of stakeholders to the framework. The result of this was a 7-gap framework that implies, for instance, that closing the gap between a developer's mental model and the AI model can both facilitate the developers understanding and acceptance of the model, but also improve the model, in cases where the decision quality of the developer's mental model is superior to the decision quality of the AI model.

[1] I changed the term "manager's" to "decision maker's" and "DSS model" to "AI model" based on the above-introduced definitions

Gregor and Benbasat (1999) proposed that users will use explanations when they observe a failure or an anomaly in the output of an AI system. Moreover, they proposed, based on Gregor (1996), that users will use explanation when they want to achieve "long-term learning" that goes beyond the mere implementation of the system generated recommendations. Here, they argued that novices would rather use explanations for learning (based on Mao 1995), while experts will use explanations for "resolving anomalies (disagreement) and for verification." (p. 514, based on Mao 1995 and Ye 1991) Also, they proposed, based on Everett (1995), Gregor (1996), and Mao (1995), that users will use explanations when they lack the knowledge needed for a problem-solving task. Concerning the complexity of explanations, Gregor and Benbasat proposed that "[e]xplanations that require less cognitive effort to access and assimilate will be used more and will be more effective with respect to performance, learning or user perceptions" (p. 514). This, they argued, based on Everett (1995) and Moffit (1989), especially affects explanations that are "always present" and "automatic," but also explanations that are "case-specific rather than generic" (based on Berry and Broadbent 1987, and Dhaliwal 1993). Finally, based on Toulmin's model of argumentation (1984 and 1958) and empirical studies by Everett (1994), and Ye (1991), they suggested that justificatory explanations, which, for instance, justify the process of how the AI system transformed data into insights, are more favorable regarding user perception and acceptance than other types of explanations.

Based on the discussion above, I theorized that a lack of AI system interpretability due to their often black-boxed and intransparent nature leads to the inability for humans to control and learn from AI, and a perception-reality gap between what a system potentially can, and what user believes it can. This, then, leads to unintended consequences, underutilization of the AI system, and algorithm aversion (lack of acceptance or rejection of the AI system; see Figure 4).

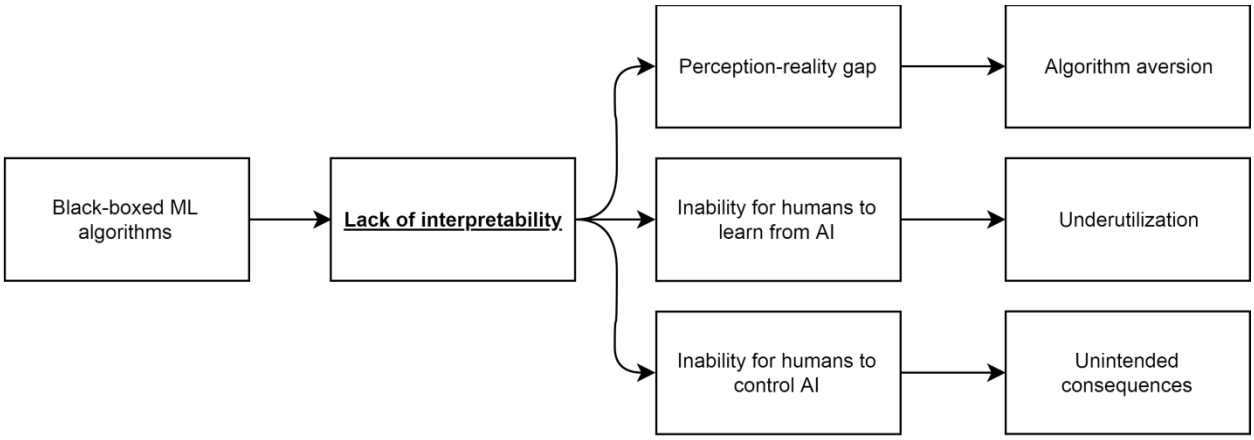

Figure 4. Lack of interpretability its cause and its consequences

Martens and Provost (2014), added to the arguments by Gregor and Benbasat as they further differentiate justificatory explanations into global and local explanations. Global explanations, they argued, are equivalent to the model parameters of linear models. However, more complex black-boxed algorithms such as neural networks (Bishop 1995) lack such built-in global explainability features. For these cases, researchers have developed decision tree-based methods that simulate the behavior of trained algorithms closely to extract rules that explain how particular variables generally contribute to predictions (Craven and Shavlik 1996; Martens et al. 2007). More recently, researchers have developed local explanation methods that help to explain the contribution of variables to a particular prediction of a machine learning model on an instance level via numerical scores (Baehrens et al. 2010; Scott M. Lundberg et al. 2018; Lundberg and S.-I. Lee 2017; Ribeiro et al. 2016a; Robnik-Šikonja and Kononenko 2008; Štrumbelj et al. 2009; Štrumbelj and Kononenko 2010).

I discussed why explanations are important in addressing the Interpretability challenge, but what are explanations actually? According to Lewis (1986), one can explain an event by providing historical information about its causes. However, as Lipton (1990) noted, "causal histories are long and wide […] The big bang, [for instance], is part of the causal history of every event, but explains only a few" (p. 249). So which causes explain an event and which do not? According to Lipton: "explanation is 'interest-relative, and […] we can analyse some of this relativity with a contrastive analysis of the phenomenon to be explained'. What gets explained is not

simply 'Why this?', but 'Why this rather than that?'" (p. 249 citing Van Fraassen 1980 and Garfinkel 1981). Based on this, Lipton introduces the concepts of "fact" and "foil." Here, a fact is an actually observed event that one tries to explain, while a foil is a contrastive counterfactual event: "We may not [simply] ask why the leaves turn yellow in November […], but only, for example […] why they turn yellow in November [, the fact,] rather than turning blue [, the foil]" (p. 249). But how does one ask a good contrastive question, and most importantly, how does one select a good foil to contrast with a fact? According to Lipton (as cited in Miller 2019, p. 9):

> [the] central requirement for a sensible contrastive question is that the fact and the foil have a largely similar history, against which the differences stand out. When the histories are disparate, we do not know where to begin to answer the question.

Based on this, Miller (2019) reasoned that to determine a foil for a sensible contrastive question, "people could use the similarity of the history of facts and possible foils" (p. 9).

Lipton (1990) proposed that the selection of explanatory causes for contrastive explanations (why facts and not foil?) is interest-relative. This selection process can be seen as "inference to the best explanation" (Harman 1965). Josephson and Josephson (1996, pp. 1–2) explicate this abductive inference process as follows:

D is a collection of data (facts, observations, givens).
H explains D (would, if true, explain D).
No other hypothesis can explain D as well as H does.
Therefore, H is probably true

Referring to the above-introduced decision making literature, one can argue that humans satisfice (Simon, 1955) when generating and selecting candidate explanations by only taking information into account that is available and easy to process at the time of explanation.

2.3.2 Accountability and Management Related AI Challenges

In psychology, researchers have focused on the effects that especially different forms of accountability have on decision making and employees. Accountability can be broadly defined as

the need to validate one's own viewpoints and acts towards others (e.g., Simonson et al. 1992). Research suggests that accountability leads to a more analytical decision making process and increased investments of time and effort (e.g., De Dreu et al. 2006; McAllister et al. 1979). Tetlock (1985; 1985; 1983) showed the debiasing effects of accountability on human judgment, e.g., by alleviating the impact of first-impressions bias (see also Lerner and Tetlock 1999; Tetlock et al. 1989). However, Tetlock and Lerner (1999) argue such effects should only occur under certain conditions:

> Accountability attenuated bias on tasks to the extent that (a) suboptimal performance resulted from lack of self-critical attention to the judgmental process and (b) improvement required no special training in formal decision rules, only greater attention to the information provided. (p. 263)

One can argue that accountability forces decision-makers to use their more analytical, rational, conscious and resource-intensive System 2 Thinking, instead of the intuitive, quick, and heuristics-based System 1 Thinking that is prone to many cognitive biases (see Kahneman, 2003).

Simonson and Staw (1992) differentiate between process accountability and outcome accountability. Under process accountability, managers have to justify their decision making processes, while under outcome accountability, they have to justify their outcomes (Simonson and Staw, 1992). They showed that process accountability reduced the commitment to initially selected decision alternatives by increasing self-critical and decreasing self-defending behaviors, while outcome accountability showed opposite effects:

> When one is accountable for outcomes, the need for justification may be heightened along with any increase in decisional vigilance. However, with accountability for process, individuals who use proper decision strategies and who thoroughly evaluate the available alternatives before reaching a decision should be favorably evaluated regardless of the decision's outcome. Thus, one could expect accountability for process to be a superior deescalation technique to accountability for outcomes. (p. 421)

According to the conflict theory by Janis and Mann (1977), situations in which the success of an important decision is highly uncertain create high levels of stress for decision-makers. Mano (1992) showed that stress could lead to simpler and more extreme decisions. Based on this, Siegel-Jacobs and Yates (1996) argued that outcome accountability could create such stressful situations, which, therefore, could affect decision quality negatively. In the same study, they showed that process accountability could increase the predictive accuracy of decision-makers by processing more of the available relevant information, while outcome accountability can decrease predictive accuracy.

I approached to understand the causes of the challenges of inflated management expectations and managing AI projects like traditional IT projects. Here I argue that the AI knowledge asymmetries between data scientists and managers and a general AI hype cause these challenges.

In recent years, many AI-related breakthroughs were made (Perrault et al. 2019). As such breakthroughs are covered broadly in the media, e.g., self-driving cars, or the first AI systems winning in games that were thought to be impossible to win for them like Jeopardy or GO, they create a hype that, in turn, leads to an increase in expectations and investments. This means that managers that believe the hype have high expectations towards AI outcomes but low-understanding of their true potentials and especially about how much effort it takes for data scientists to create even intermediate AI outcomes. As a result of this, they treat AI projects just like IT projects and expect plannable outcomes. However, many AI projects are highly complex and still, in many cases, fail (Silver 2012). Based on this, one can argue that AI projects have a high outcome uncertainty. In such situations, research strongly suggests to evaluate employees not based on their outcomes, but rather on their inputs and processes to achieve such outcomes, as otherwise, they may be blamed for something that was fully out of their control (Tetlock et al. 2013). For managers, high expectations in AI projects paired with the high outcome uncertainty of AI projects should increase the potential for severe cognitive dissonance significantly (see Festinger 1957). For data scientists, this should increase their stress levels even more, as they should be more or less aware of the true potentials of AI and the effort it takes to exploit them.

Sharma (1997) argues that in such situations, when a layperson (in this case, an AI hyped manager) and an expert (in this case, a data scientist) interact, knowledge asymmetries occur that have several negative effects. According to Sharma, this knowledge asymmetry leads to the opportunistic behaviors of experts. In such cases, Sharma (1997) argues that firms need to install behavioral control mechanisms in contrast to evaluating experts based on their outcomes only. Here, Sharma proposes that such behavioral controls require superordinate supervisors. However, Sharma focuses on cases in which it is comparatively easy for the expert to create an outcome, and, therefore, also easier to find superordinate supervisors. In the AI case, where finding superordinates that evaluate the work of data scientists should be difficult, it could lead to a vicious circle in which data scientists initially trigger the hyped manager even further to get valuable resources, and support for their projects, knowing that the success probability is low, which should have negative effects on the firm, and in the long run on the data scientist that cannot deliver. Alternatively, it could lead to a scenario where the stress level of data scientists increases drastically, both because of the general outcome uncertainty involved in AI projects and because of a manager that will potentially be very unsatisfied due to the expectation that with AI, one can create high value easily.

Based on this discussion, I theorized that the AI hype and knowledge asymmetries lead to inflated management expectations and managing AI projects like traditional IT projects. Furthermore, I theorized that this leads to a situation in which data scientists are not made accountable for their actions at all, or they are solely made accountable for the outcomes they deliver, which, as discussed above can, in the long run, be bad both for managers, the data scientists, and the firm (see Figure 5).

Nevertheless, this implies a potential solution for these challenges too. As data scientists could use their expertise (knowledge asymmetries) and, e.g., AI interpretability methods to show managers a more realistic picture of what AI actually is and can do. At the same time, they would enable managers to better understand the effort and processes that go into the creation of an AI system, which would consequently allow managers to make data scientists process instead of outcome accountable only.

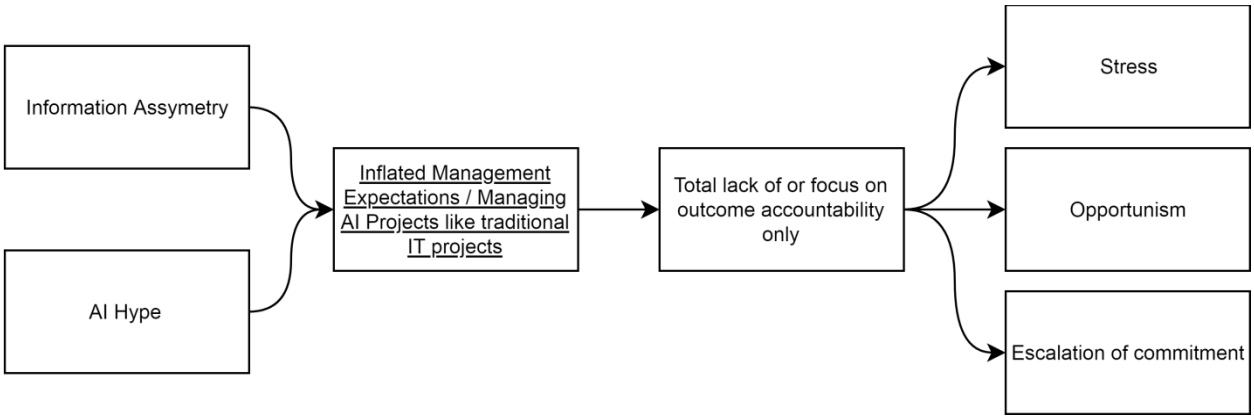

Figure 5. Management-related AI challenges and some of its consequences

2.3.3 Causality and Causal Inference

One of the key challenges that we identified in (see Section 5.1) our AI value creation study was causality. And, in particular, the confusion of associational questions with causal questions (see Leek and Peng 2015). When answering causal questions, one wants to identify and quantify causal relationships between a cause (treatment) and an effect (outcome). However, the interpretation of such cause and effect relations can lead to biased conclusions due to spurious associations. According to Hernan et al. (2002):

> Intuitively, two variables E and D will be statistically associated if one is a cause of the other (e.g., smoking and lung cancer), if they share a common cause (e.g., yellow fingers and lung cancer share smoking as a common cause), or both. If E precedes D, the overall association between E and D will have two components: a spurious one that is due to the sharing of common causes [, confounding,] and another due to the causal effect of E on D. (p. 176 – 177)

The goal of causal analysis is "to eliminate a spurious association," and ways to do this are "to adjust, stratify, or condition on the common cause; for example, we would find no association between yellow fingers and lung cancer among nonsmokers." (Hernan, 2002, p. 177) Following this, "[c]onfounders are variables that when stratified on or adjusted for will eliminate (or diminish) the spurious component of the association between exposure and disease." (Hernan, 2002, p. 177) A researcher's assumptions about confounders can be represented as causal dia-

grams in the form of directed acyclic graphs (DAGs; Greenland et al. 1999; Hernán et al. 2002; Pearl 1995, 2009) as depicted in Figure 6. Based on the example by Hernan et al. (2002), here, E could represent a variable that measures whether subjects of a study had yellow fingers, D could measure whether they had developed lung cancer, while C could be a common cause (confounder) smoking of both E and D. If one is interested in the true causal effect of E on D, one has to eliminate the spurious association that C introduces. A way to do this would be to condition on C. This way, for instance, one could look at the effect that yellow fingers have on lung cancer among nonsmokers only. In this case, as mentioned above, one would not find a statistical association, which indicates that no such causal relationship exists. Hernan et al. (2002) emphasized the importance of researchers' knowledge in causal analyses: "We wish to emphasize that causal inference from observational data requires prior causal assumptions or beliefs, which must be derived from subject-matter knowledge, not from statistical associations detected in the data." (p. 181)

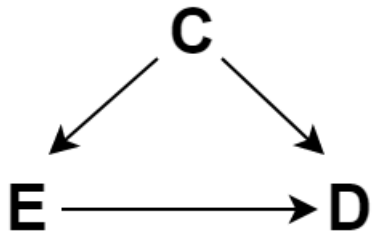

Figure 6. Example of confounding

Other causes of the Causality challenge are selection, measurement bias, and multi-collinearity. Selection bias occurs, for instance, when one wants to make inferences for a whole population of subjects based on an experimental study in which subjects were not selected randomly. Hernán et al. (2004, p. 615) provide an alternative definition: 'A structural classification of bias distinguishes between biases resulting from conditioning on common effects ("selection bias") and those resulting from the existence of common causes of exposure and outcome ("confounding").' Conditioning on common effects, could for instance happen, if one wants to estimate the effects of a price change (treatment) on materials (subjects), but looks, for instance, only at the materials that had an increase in sales volume (effect).

Measurement bias, on the other hand, simply relates to the issues that occur either due to low-quality data or because one wants to answer questions that the data does not permit. A typical example of measurement bias is the use of the Body Mass Index to measure obesity (Hernán and Cole 2009). As Hernán and Cole (2009), state: "Causal inferences about the effect of an exposure on an outcome may be biased by errors in the measurement of either the exposure or the outcome." (p. 959)

Another threat to causal analyses that causal diagrams help to eliminate is collinearity or multicollinearity (Schisterman et al. 2017). It occurs when two or more independent variables in a model are highly correlated. It can, for instance, occur when a third variable (e.g., C) mediates the effect of a treatment (e.g., E) on an outcome (e.g., D; see Schisterman et al. 2017 and Figure 7).

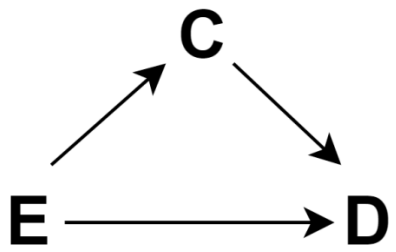

Figure 7. Example of collinearity introduced by a mediator variable

Collinearity and multicollinearity are less relevant when building predictive models but very important when doing causal analyses:

> Multicollinearity is not a problem unless either (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to Y, without the influence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict. (Wheelwright et al. 1998, p. 288)

Based on this discussion, I theorized that confounding, selection bias, measurement bias, and (multi-)collinearity are all causes of the Causality challenge, which, in turn, leads to biased explanatory knowledge, and eventually low decision quality (see Figure 8).

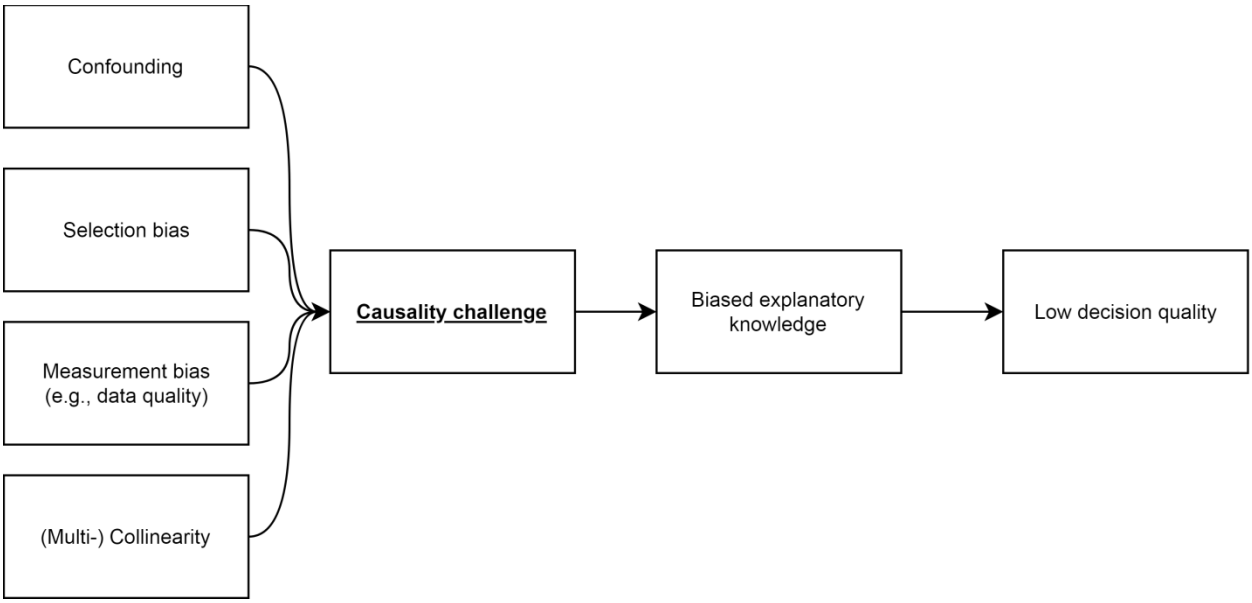

Figure 8. Causes and consequences of the Causality challenge

The common perception in quantitative research is that randomized controlled trials (RCTs) are the only approach that can really answer causal questions in an unbiased way (Cochrane 1972). Based on this, RCTs have been called the "gold standard" of causal analysis (Cartwright 2007). When perfectly set up, RCTs should remove all confounding (Cartwright 2007, 2009). Such RCTs assure a high internal validity (the degree to which the analysis can answer the question asked), but often lack external validity (generalizability to other contexts; Rothwell 2006).

Designing and conducting internally valid RCTs is a comparatively complex task, e.g., one has to randomly assign subjects to different groups, exposing them to different treatments, and measuring the difference of such treatments on an outcome variable. Schulz (1995) could show that a significant number of RCT-based studies used inadequate approaches, which resulted in biased treatment effect estimates. Due to such issues, Cartwright (2007) argued that "[t]here is no gold standard."

Nowadays, researchers and organizations often have large amounts of observational data at hand. While doing causal analyses on such data may be preferable in terms of external validity, they introduce some additional challenges for internally valid analyzes (Kleinberg and Hripcsak 2011). A causal inference approach, which is somewhat in-between the spectrum of

RCTs (high complexity, high internal, low external validity) and purely observational analyzes (low complexity, low internal, and high external validity), are quasi-experimental analyzes (medium complexity, medium internal, medium external validity; Price et al. 2015). In quasi-experimental analyzes (Cook et al. 1979; Shadish et al. 2002), just like in RCTs, one does manipulate an independent treatment variable. However, one does not assign subjects to groups randomly. Moreover, since one manipulates the treatment variable before one studies its effects on the outcome, the direction of a causal effect in quasi-experimental analyzes is already predefined (in contrast to purely correlational studies that only indicate an association between two variables but do not indicate which variable is the cause and which is the effect). Therefore, quasi-experiments solve the directionality problem of correlational analyzes but still suffer from confounding due to the non-random assignment of subjects (Price et al., 2015).

One type of quasi-experiments are difference-in-differences (DID) analyzes. In DID analyzes, such as Ashenfelter and Card (1985) and Card (1990), at first, one calculates the difference between the averages of an outcome variable (e.g., sales of materials in euro) for a treatment group (received treatment) and a control group (did not receive the treatment) in the pre-treatment period. Afterward, one does the same calculation for the post-treatment period. Finally, one takes the difference between those differences to get the treatment effect. Another approach called propensity score matching (Abadie and Imbens 2006; Rosenbaum and Rubin 1983) uses data of the pre-intervention treatment, pre-intervention control outcomes, and additional pre-intervention covariates to calculate scores based on which one matches each treatment subject with one or more control subjects. Another approach is the synthetic control method (Abadie et al. 2010; Abadie and Gardeazabal 2003). In synthetic control methods, one constructs a synthetic control group as a weighted combination of control subjects that matches the treated subject closely in terms of pre-intervention period covariates and outcomes. The post-intervention values of this synthetic control group simulate the counterfactual outcome of the treated subject that one could have expected to observe had the treatment not occurred. So while in propensity score matching, one finds one or more closely matching control subjects for

a treated subject, in synthetic control methods, one constructs a closely matching control group for a treated subject.

The foundation of causal inference is the idea that one wants to compare the observed outcome of a subject after treatment with the outcome that would have been observed had the subject not been treated (Rubin 1974). While under the condition that time traveling is not an option, one cannot observe the second outcome, the above-described methods all try at least to simulate such counterfactual cases.

Following this, Hal Varian (2014, 2016), chief economist at Google, argued that constructing counterfactuals, in its essence is a predictive task: "The better predictive model you have for the counterfactual, the better you will be able to estimate the causal effect, a rule that is true for both pure experiments [, quasi-experiments,] and natural experiments." (2014, p. 21) Moreover, he emphasized that while in scientific experiments and quasi-experiments, researchers focus on how a treatment (e.g., a drug) affects a subject (e.g., a patient), in business applications like in marketing, 'the primary interest is often in how the treatment of the subjects affects the "experimenter."' Moreover, he implied that in such settings, practitioners are less interested in how exactly an intervention, e.g., on ad spend affects an outcome, e.g., webpage visits: "whether the increase is attributable to ad clicks, search clicks, or direct navigation may be of secondary importance." (p. 7312) Based on this, he argued, that predictability may be a more important characteristic when selecting variables for simulating counterfactuals.

Following this, Varian (2016) proposes a conceptually simple approach to causal inference where one, first, manipulates a treatment variable such as ad spent for a certain period of time. The observed outcome, e.g., webpage visits, for the post-intervention period is then compared with the counterfactual case in which the treatment (ad spent) was not changed. He added: "[h]owever, where does the counterfactual come from? Answer: it is a predictive model developed using data from before the experiment was run" (p. 7312). Brodersen et al. (2015), Varian's colleagues at Google, developed with causal impact analysis such a prediction focused causal inference approach. In causal impact analysis, one defines a point of intervention, e.g., when the

ad spent was increased. Then one uses time-series data of the pre-intervention period to estimate a counterfactual via a Bayesian structural time-series model (BSTS; Scott and Varian 2014) for the post-intervention period. To calculate the causal impact (treatment effect), one takes the difference between the observed outcomes and the counterfactual outcomes in the post-intervention period. Here, one can basically use all variables that are highly correlated with the outcome in the pre-intervention period but were not directly affected by the intervention (treatment). An example of this can be lagged variables of the outcome itself, seasonality, holidays, working days, but also economic indices. Moreover, Brodersen et al. (2015) encourage the use of potential control groups as well. Such control groups could be, in the ad spent case, users in other regional markets, whose behavior was similar to the treated users before the intervention, e.g., because the ad spent was similar. Varian (2014) described this general approach as follows:

> This procedure does not use a control group in the conventional sense. Rather it uses a general time series model based on trend extrapolation, seasonal effects, and relevant covariates to forecast what would have happened without the ad campaign. A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard. To see this, suppose that you run an ad campaign in 100 cities and retain 100 cities as a control. After the experiment is over, you discover the weather was dramatically different across the cities in the study. Should you add weather as a predictor of the counterfactual? Of course! If weather affects sales (which it does), then you will get a more accurate prediction of the counterfactual and thus a better estimate of the causal effect of advertising. (p. 24)

Moreover, to reduce the complexity of having too many variables, the approach automatically selects variables with the help of spike and slab priors (Scott and Varian 2013).

3 The B2B Aftersales Context at MAN Energy Solutions

MAN Energy Solutions is an original equipment manufacturer of 2- and 4-stroke diesel engines, for ocean-going vessels and power plants, but also produces gas engines, dual-fuel engines, and turbomachines. They also offer power-generating 4-stroke engines, as well as propulsion solutions and turbochargers. Moreover, their global aftersales organization offers original spare parts as well as ad hoc and contract-based operating and maintenance services through a worldwide network of local sales companies. We conducted the ADR studies in MAN Energy Solutions' aftersales service organization that operates in a complex and dynamic market, which is heavily dependent on the number of newly built ocean-going vessels (Danish Ship Finance 2018). At this moment, the new sales market for ship engines stagnated, as current vessel numbers were high. Because of this, the aftersales service business became more critical for MAN Energy Solutions.

This shift of focus made some significant changes to the company's aftersales processes necessary. Usually, they approached customers via reactive mechanisms and intensive key-account management. To better exploit the potential of the aftersales market, however, MAN Energy Solutions wanted to improve its traditional aftersales approach. They started, for instance, a digitalization initiative, of which a cornerstone was creating more proactive sales processes and services based on a thorough understanding of customer needs.

While their market for ship new building was very transparent, their aftersales market was less transparent. In the market for ship new building, most large shipyards were well known, as were the main competitors. Moreover, one must register new ships with the international maritime organization, which requires detailed information on the type of ship, ownership structure, and type of machinery. Moreover, MAN Energy Solutions collaborated with licensing partners that manufactured most of their original equipment and were, therefore, valuable trading partners.

In the aftermarket, it was not always clear for MAN Energy Solutions, who the competitors were, because not just original equipment manufacturers could service their equipment but also third-party companies, which, however, could usually not assure the same quality and security

that MAN Energy Solutions could provide. The same licensees that manufacture MAN Energy Solutions' equipment were amongst the toughest competitors in their aftermarket as they could utilize regional networks with MAN Energy Solutions' customers. Moreover, the aftermarket was uncertain since they only made a limited number of long-term service agreements with their customers. Thus, it was, for instance, unclear when a customer relationship started, when it ended, and when one could expect the next sale. Researchers call this a "noncontractual" setting (Fader and Hardie 2009). Such an aftersales setting usually also shows heterogeneous and "lumpy" demand patterns (Bartezzaghi et al. 1999; Bartezzaghi and Kalchschmidt 2011).

Sales professionals at MAN Energy Solutions recommended how many running hours of the engine a spare part could last until it needs replacement. The problem was that they could only give some general advice to customers in the form of manuals, but they could only seldomly check the actual running hours of an installed engine during maintenance jobs. In reaction to this inaccessibility, MAN Energy Solutions started an initiative that aimed to access "live" engine data via smart assets, interfaces, and networks. Moreover, as this initiative was still quite new, they approached to estimate an engine's running hours via satellite data from the automatic vessel identification system. This PhD project was another initiative to utilize data and advanced technology to enable more engine and customer lifecycle-based aftersales approaches.

4 Methodology

The following chapter describes first the action design research method that I used during this PhD project (Section 4.1). Afterward, Section 4.2 discusses design principles and how they were formulated during the ADR studies. Then Section 4.3 discusses what differentiates consulting, from design science research, and how design principles relate to design theory. And Section 4.4 describes the qualitative research method with which my co-authors and I conducted the AI Value Creation Study.

4.1 Action Design Research

I framed this PhD project as an action design research (ADR; Sein et al. 2011) program that consists of three ADR sub-studies and one qualitative interview study that informs the ADR program. In each ADR study, we designed and implemented artifacts that belong to the overall class of AI system for customer-centric B2B aftersales decision support.

Sein et al. (2011) introduced the ADR method as an alternative to the design science research (DSR) methods by Hevner et al. (2004) and March and Smith (March and Smith 1995) that according to them focus too much on the sequential design and evaluation of IT artifacts, while largely neglecting socio-technical context. Based on the definition by Orlikowski and Iacono (2001), Sein et al. (2011) argue that one should understand IT artifacts as representations of the socio-technical relations that went into its construction:

> This definition reflects a "technology as structure" view of the ensemble artifact, where structures of the organizational domain are inscribed into the artifact during its development and use […]. It accommodates designers' building and organizational stakeholders' shaping in a single definition, thereby softening the sharp distinction between development and use assumed in dominant DR thinking. (Sein et al. 2011, p. 38 citing Orlikowski and Iacono 2001)

Based on these arguments, they propose a method that consists of aspects of both artifact-focused DSR and context focused action research (AR; Baskerville and Wood-Harper 1996; Susman and Evered 1978). However, ADR is more than a combination of DSR and AR. In ADR, one starts with designing an initial theory ingrained solution in a DSR fashion but then intervenes with it into an organizational context (Sein et al. 2011), in reaction to the organizational shaping of the artifact, one continuously circles between design, intervention, evaluation, redesign. A traditional DSR study, on the other hand, usually stops at the conceptual design of an artifact and attempts to demonstrate and evaluate its utility with methods such as focus groups. Based on such evaluation, one then either starts a new conceptual design iteration or when it's the evaluation was positive publish the results. While such forms of evaluation are possible in ADR too, the continuous evaluation in ADR is much more dynamic. When implementing a

conceptually designed prototype of an information system in a real organizational context, for instance, it could turn out that the data quality is much worse than assumed, or that users do not accept and consequently use the system for reasons that the designers did not anticipate. Such evaluations require immediate adjustments to the artifact design; otherwise, one might lose valuable momentum and, eventually, organizational support. Most likely, one would not get such feedback from demonstrating a prototype during a focus group session.

In concrete, the ADR method consists of four general stages and seven principles (Sein et al. 2011; see Figure 9).

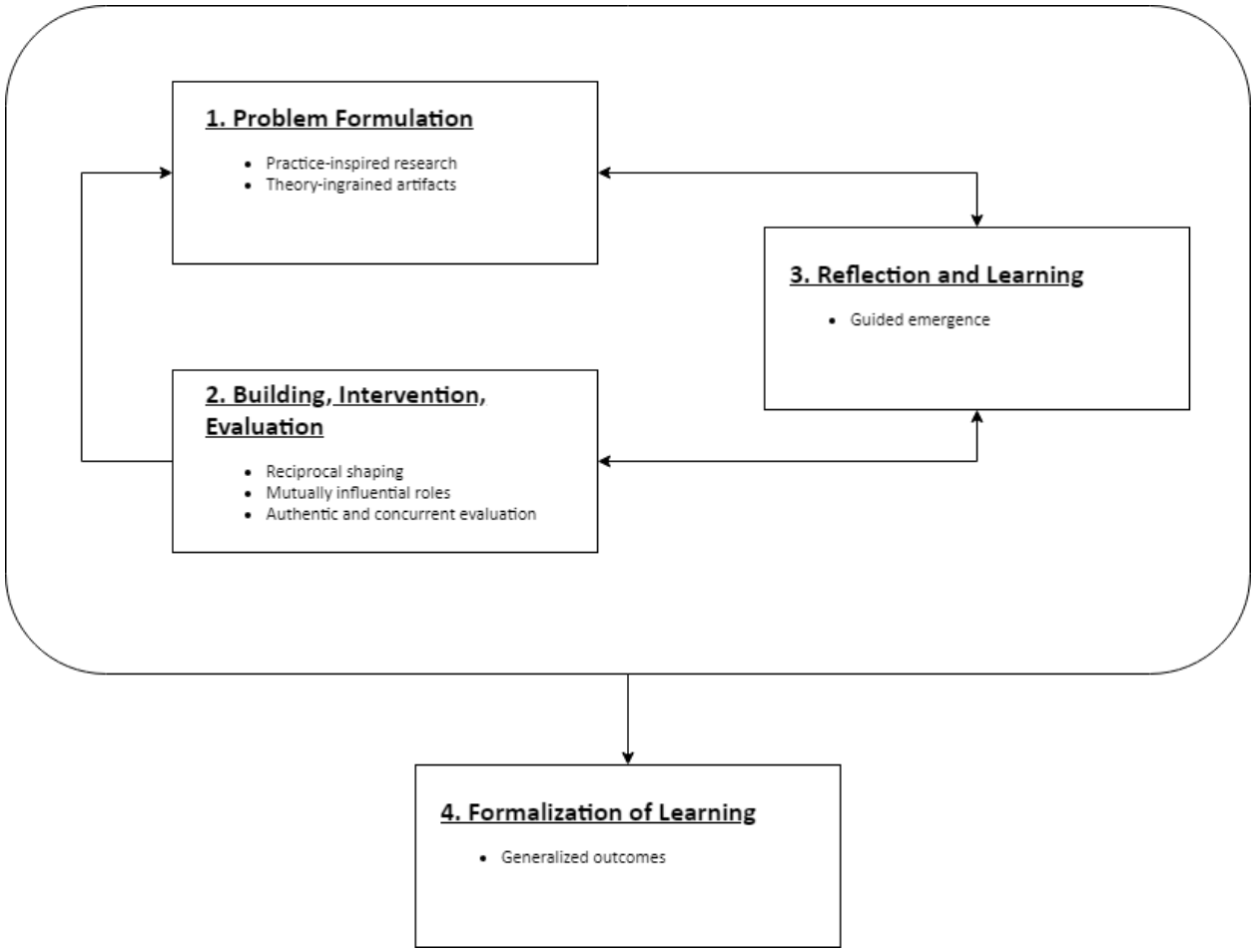

Figure 9. ADR method (Sein et al. 2011)

The principles of *practice-inspired research* and *theory-ingrained artifact* are the fundaments of the problem formulation stage. The *practice-inspired research* principle is about perceiving practical field problems as opportunities for knowledge creation:

> The intent of the ADR team should not be to solve the problem per se as a software engineer or a consultant might. Neither should it be only to intervene within the organizational context of the problem. Instead, the action design researcher should generate knowledge that can be applied to the class of problems that the specific problem exemplifies. (Sein et al. 2011, p. 40)

The *theory ingrained artifact* principle, in turn, emphasizes that ADR artifacts are theory-informed. Here, theory refers to Gregor's (2006) five types of theory. In particular, ADR researchers can use theory "to structure the problem (…), to identify solution possibilities (…), and to guide the design" (Sein et al. 2011, p. 41). They note, however, that "this act of inscribing […] results in only the initial design of the theory-ingrained artifact. It is then subjected to the organizational practice, providing the basis for cycles of intervention, evaluation, and further reshaping." (p. 41) Typical tasks in the problem formulation stage are (p. 41):

(1) Identify and conceptualize the research opportunity
(2) Formulate initial research questions
(3) Cast the problem as an instance of a class of problems
(4) Identify contributing theoretical bases and prior technology advances
(5) Secure long-term organizational commitment
(6) Set up roles and responsibilities

I have approached the problem formulation stages slightly different for each ADR study; however, on the more general level of the ADR program, they all follow a coherent approach. I have identified the field problem of costly and undifferentiated B2B aftersales service strategies due to a lack of customer-centricity. This allows low-cost competitors to threaten the potential gains of servitization and aftersales service focus (Section 1). Also, researchers argue for the potentials of technology and particularly AI, but that guidance for practitioners in B2B contexts is

sparse. I conceptualized this field problem as an opportunity to create knowledge about *how to design AI systems that support customer-centric aftersales processes and strategies.* While for the ADR sub-studies, I had to abstract and cast the problem as an instance of a class of problems, the overall ADR program was already on that abstract level. In the next step, I identified with AI value-creation and industrial marketing literature contributing theoretical bases and particular with probabilistic models for customer-lifetime value estimations prior technological advances. Moreover, during the literature review in *ADR Study 1*, a knowledge gap became apparent. I addressed this gap with an *AI value creation study* that helped to structure the problem and to identify solution possibilities for the remaining two ADR studies. Moreover, I secured long-term organizational commitment with a research collaboration agreement, similar to a researcher-client agreement in AR (Davison et al. 2004), that the IT University of Copenhagen and MAN Energy Solutions set-up as part of the overall industrial PhD project. Finally, I set-up roles and responsibilities (see Figure 10). In general, during the design and implementation of all three AI systems (ADR studies 1-3), I was both the lead researcher and lead developer. While practitioners at MAN Energy Solutions were partly involved in the development and organizational support, they mostly contributed to domain knowledge and end-user feedback. Moreover, my supervisor and main co-author, Oliver Müller, was most of the time in the role of an external researcher that could help to structure and abstract the situated problems and solution designs at MAN Energy Solutions.

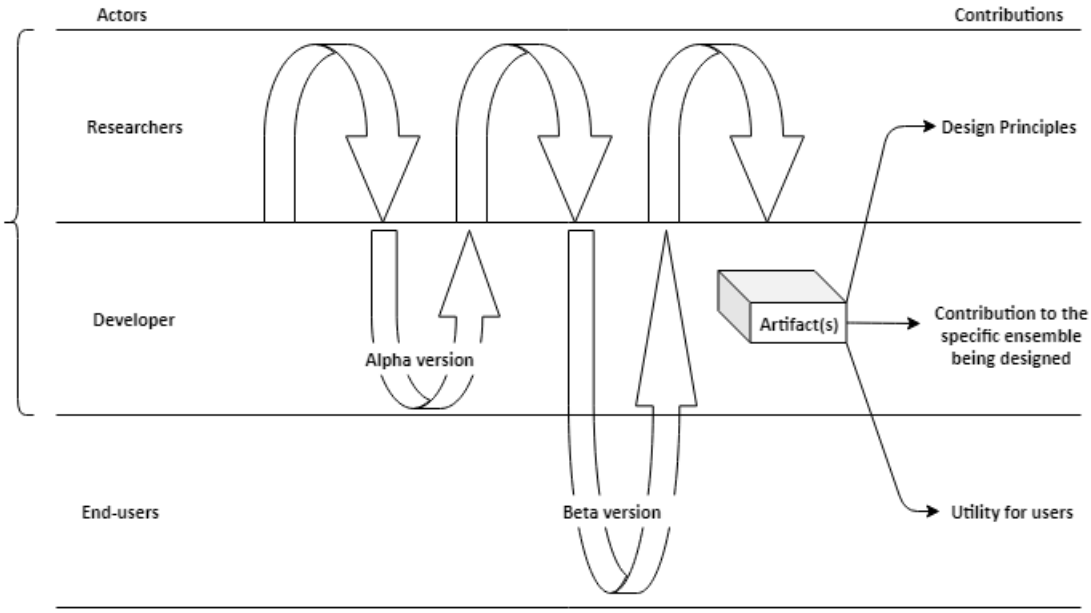

Figure 10. ADR roles (Sein et al. 2011)

The principles of *reciprocal shaping*, *mutually influential roles*, and *authentic and concurrent evaluation* are the fundaments of the *building, intervention, and evaluation* stage (BIE; Sein et al. 2011).

The *reciprocal shaping* principle emphasizes the close interconnection of how an artifact shapes its environment after the intervention and how the environment in the following shapes the artifact. For example, "the ADR team may use its chosen design constructs to shape its interpretation of the organizational environment, use this increasing understanding of the organizational environment to influence the selection of design constructs, and/or interleave the two" (Sein et al. 2011, p. 43).

The *mutually influential roles* principle emphasizes the shared learning that occurs when researchers bring their theoretical and technical knowledge, while the practitioners contribute with subject-matter knowledge and practice-based hypotheses derived from organizational work experiences (Sein et al. 2011). The researchers' and practitioners' contributions can both complement but also contest each other (Mathiassen 2002).

The *authentic and concurrent evaluation* principle is the crucial ADR differentiator from the traditional stage-gate like DSR approach (see Hevner et al. 2004; Sein et al. 2011). In ADR, the evalua-

tion of the alpha versions of artifacts (see Figure 11) is usually formative, while the evaluation of beta versions is summative, e.g., evaluating how effective the artifact was in solving the field problem (Sein et al. 2011). However, due to the "emerging" nature of ADR artifacts, "authenticity is a more important ingredient for ADR than controlled settings" (p. 44). Remenyi and Sherwood-Smith (1999) describe formative evaluations as follows:

> [w]hen systems analysts take their first glimpse at the information system supporting some departments activity they have very little knowledge of what the systems are and how the system work, but at the end of the first day they will have documented their analysis and the next day will present it to the department head who will jointly with the analyst evaluate the documented results. The analyst will express his or her impression of the system; the department head will endeavour to understand the documented impression, both will have some issues to clarify and the understanding and perception of both parties will have evolved as a result of the exercise. The analysis and design process continues in this way and design decisions continue to be supported by this formative evaluation process. The formative evaluation process is further influenced by the objectives of the information systems development project. The systems analyst and head of department did not meet by chance. They were given terms of reference and systems objectives, and part of the formative evaluation process for both of them is to understand these terms of reference and systems objectives and then to evaluate their progress and design in the context of the objectives. (p. 24)

Figure 11 illustrates such a formative evaluation process that both acknowledges changing perceptions of the solution, but also changing perceptions of the design goals to achieve.

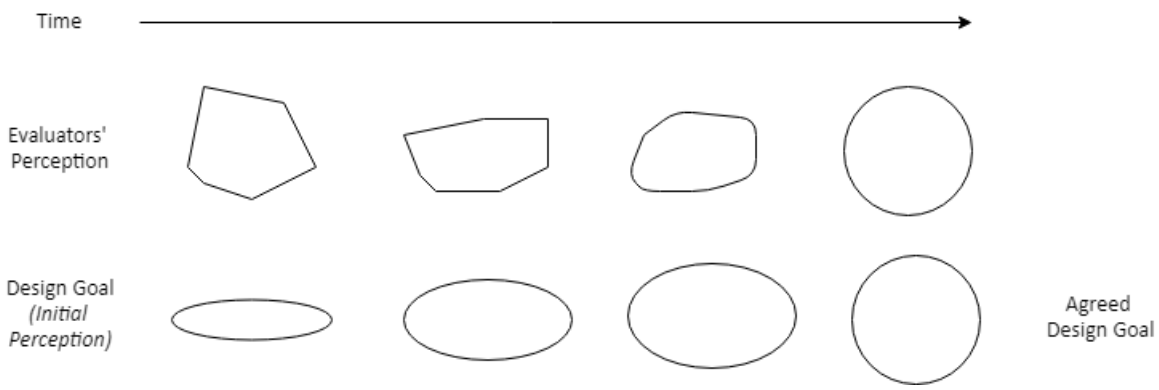

Figure 11. A Formative Evaluation Process adapted from (Remenyi and Sherwood-Smith 1999)

In the BIE stage, typical tasks are (Sein et al. 2011, p. 43):

(1) Discover initial knowledge-creation target

(2) Select or customize BIE form

(3) Execute BIE cycle(s)

(4) Assess need for additional cycles, repeat

At the beginning of each ADR sub-study, we were defining design goals and constructed an initial alpha version of an IT artifact in close collaboration with practitioners and managerial stakeholders at MAN Energy Solutions. At this stage, "[e]arly designs and alpha versions serve[d] as lightweight interventions in a limited organizational context" (Sein et al. 2011, p. 42).

In all our ADR sub-studies, we used a rather "IT-dominant BIE" form as we focused more on how the environment shaped the artifact and less on how the artifact shaped the environment ("organization-dominant BIE"; Sein et al. 2011) In an IT-dominant BIE form:

> The emerging artifact, as well as the theories ingrained in it, are continuously instantiated and repeatedly tested through organizational intervention and subjected to participating members' assumptions, expectations, and knowledge. This highly participatory process builds organizational commitment and guides the eventual design of the ensemble artifact. (Sein et al. 2011, p. 42)

Based on this, we executed the repeated BIE cycles until we evaluated our solution artifacts as satisfactory, which in concrete means that they met the agreed design goals and helped to solve the field problem.

The *reflection and learning* stage is about continuously abstracting the learnings from the situated design of IT artifact to the broader class of problems and solutions that this artifact belongs to (Sein et al. 2011). As Sein et al. (2011) put it: "[c]onscious reflection on the problem framing, the theories chosen, and the emerging ensemble is critical to ensure that contributions to knowledge are identified" (p. 44). The principle of *guided emergence* is the fundament of this stage. The principle emphasizes that the eventual ensemble artifact will not only consist of traits of the initial theory-inscribed solution design but also of the interactions with the organization and the *authentic and concurrent evaluation.* When comparing the initial solution design and the final design of the ensemble artifact, one should expect "substantial changes" or "mutations" rather than "trivial fixes" (Sein et al. 2011, citing Walls et al. 1992 and Gregor and Iivari 2007). According to Sein et al. (2011), "[a]nticipated as well as unanticipated consequences prompt these refinements during the BIE iterations, which provide an opportunity for the ADR team to generate and evolve design principles throughout the process" (p. 44).

In the reflection and learning stage, typical tasks are (Sein et al. 2011, p. 44):

(1) Reflect on the design and redesign during the project
(2) Evaluate adherence to principles
(3) Analyze intervention results according to stated goals

In our three ADR sub-studies, we continuously reflected on the design and redesign and conceptualized initial, emerging, and the final design principles with the help of a "coding" logbook. In general, we had access to large amounts of data. First of all, we had access to MAN Energy Solutions´ transactional aftersales services systems and databases. Also, we had access to many documents, such as process descriptions, slideshows about many strategic decisions and initiatives, standards and policy documents, etc. Moreover, we could participate in many workshops and events that were related to the company's aftersales service and digital strate-

gies. Most of all, to evaluate and refine our artifacts, we conducted regular development meetings and presentations with key stakeholders such as managers and end-users. From such encounters, we wrote rich observational notes, following an approach similar to what Baskerville and Myers (2015) call "design ethnography." In contrast to anthropological ethnography: "[d]esign ethnography is where the researcher goes beyond observation and actively engages with people in the field" (Baskerville and Myers 2015, p. 1).

Moreover, we evaluated adherence to principles. However, we focused less on whether a selected theory failed or succeeded in predicting anticipated consequences, but rather on which design principles contributed most to fulfilling the design goals and solving the field problems. This is because our design principles are targeted at practitioners and researchers trying to design effective AI systems for customer-centric B2B aftersales decision support and not at the researchers that developed the theories which inform our design principles.

In the *formalization of learning* stage, one develops more general theoretical concepts, based on the situated learnings from designing an IT ensemble artifact, that address a whole class of problems and solutions. The principle of *generalized outcomes* is the basis of this stage. In particular, Sein et al. (2011, p. 44) propose three levels of generalization:

> The first level consists of casting the original problem as an instance of a class (following the foundation laid in Principle 1). The second level entails reconceptualizing the specific solution instance into a class of solutions because an ADR effort will result in a highly organization-specific solution. The third level requires reconceptualizing the learning from the specific solution instance into design principles for a class of solutions.

In this stage, the typical tasks are (Sein et al. 2011, p. 45):

(1) Abstract the learning into concepts for a class of field problems
(2) Share outcomes and assessment with practitioners
(3) Articulate outcomes as design principles
(4) Articulate learning in light of theories selected
(5) Formalize results for dissemination

In the different ADR sub-studies, we abstracted the learning to different classes of field problems, however, for all artifacts and design principles, the overall class of field problems is underutilization of servitization and B2B aftersales service potentials. We have shared our outcomes with practitioners, while some practitioners with a strong analytical focus were interested in the theoretical contributions as well, most practitioners were mostly interested in the implemented systems and their performance. Our main theoretical contribution was the development and formalization of design principles, while we, as mentioned before, focused less on testing or extending the theories that informed them.

4.2 Design Principles

Van Aken (2004) calls for "prescription-driven" rather than "description-driven" research to help managers in solving actual field problems. For her, prescription-driven research produces "management theory" based on so-called "technological rules" that are of a heuristic nature, while description-driven research develops "organization theory." She defines a technological rule as "a chunk of general knowledge, linking an intervention or artefact with a desired outcome or performance in a certain field of application" (p. 228). Such rules are closely related to design principles in IS design science research (IS-DSR; Gregor et al. 2020). While most design science researchers acknowledge the importance of design principles, some researchers do not perceive them as design theory (or parts of it). Instead, they refer to them as constructs or methods that accompany the designed artifacts (e.g., Hevner et al. 2004; March and Smith 1995). Gregor and Jones (2007), on the other hand, see design principles, and other constructs, methods, or models as "components" of theory.

Researchers develop design principles to transfer knowledge that is broader than the description of the situated implementation of IT artifacts (Gregor and Hevner 2013; Gregor and Jones 2007; Kruse et al. 2015; Seidel et al. 2018). In a design science research paper, researchers can present design principles in the discussion section or a specific design principles section. This way, readers are supported by explicitly formulating the prescriptive guidelines that they otherwise would have to deduce from the description of the artifact (Kruse et al. 2015).

Researchers should formulate design principles concisely so that designers can follow them easily (Kruse et al. 2015). Gregor and Jones (2007) argue based on Popper's theory of three worlds (Popper 2002) that both material artifacts (physically existing things), abstract artifacts (e.g., descriptions of methods, or theories) and subjective understandings of such artifacts exist. Following this, one can then develop and formalize design principles based on subjective impressions from observing and interacting with material artifacts or by deducing design principles from abstract artifacts (Gregor and Jones 2007; Kruse et al. 2015; Shirley et al. 2013). According to Kruse et al. (2015, p. 4040), "[d]esign principles interpret descriptive, explanatory, and predictive knowledge—which can be referred to as the kernel theory—into something that can be used in the practice of building purposeful IS artifacts."

Besides their usefulness in transferring prescriptive design knowledge, design principles allow generalizing design knowledge from the situated implementation of IS artifacts to a broader class of problems and solutions (Gregor and Hevner 2013; Kruse et al. 2015). In addition to this, they are a key component of more mature design theory (Gregor and Jones 2007). One can argue that design principles are what differentiates design science research from the practice of design (Iivari 2007; Kruse et al. 2015).

Kruse et al. (2015, p. 4042-4043) empirically identified three types of design principles: "action oriented design principles" that prescribe what an artifact should allow its user to do, "materiality oriented design principles" that prescribe what features an artifact should have, and "action and materiality oriented design principles" that prescribe what an artifact should allow its user to do via a particular feature. The third type, therefore, combines the two preceding types. However, Kruse et al. (2015) noted that such types of design principles sometimes suffer from imprecise and inconsistent formulations. In reaction to this imprecision, they present a formulation template for design principles. The template suggests that one should formulate design principles in terms of how the features of an artifact ("material properties") allow its users to perform particular actions in the presence of specific "boundary conditions" (Kruse et al. 2015, p. 4044).

In the three ADR sub-studies and the corresponding papers, we followed this template, however, not strictly. We formulated the material properties, for instance, less descriptively and more imperatively targeted toward a particular designer, e.g., a data scientist, and a recipient. For instance, when following the template strictly, we could have stated, "the artifact should contain traces of domain knowledge [, the material property,] to allow managers [, the recipient,] to accept the artifact [, the action]." However, we stated, "incorporate domain knowledge [, the imperative towards a designer to add a material property,] into the data-driven decision-making process [, the artifact,] to encourage acceptance [, the action,] by managers [, the recipient]." Moreover, instead of focusing on the actions that a material property enables, we sometimes focused on the effects or outcomes that the features shall cause. For instance: "schedule regular management presentations [, the imperative towards a designer to add a material property,] to increase data scientists' [, the recipients,] need for justification [, the effect or outcome]."

4.3 Consulting, Design Principles, and Design Theory

Information systems research (IS) is about effectively designing, delivering, and using information technology, but also about the evaluation of its organizational and societal impact (Avison and Fitzgerald 1995). IS uniquely differentiates itself from other fields in its focus on artifact design and uses in socio-technical systems (Gregor 2002). It is, therefore, a discipline that concerns both the creation of knowledge about physical artifacts and behavioral phenomena (Gregor 2002). Here, the term artifact generally refers to human-made, and thus, artificial things (Simon 2019). While IS researchers have traditionally focused on developing descriptive, explanatory, and predictive theories about IT-based artifacts (Gregor 2006), an increasing amount of IS research shifts the focus towards the development of prescriptive theory about how to design such artifacts (Gregor and Jones 2007).

Such design theory is an alternative to the traditional descriptive, explanatory, and predictive theory types (Gregor 2006). Design theory is, according to Gregor (2002), "a normative or prescriptive type of theory – it gives guidelines or principles that can be followed in practice." This prescriptive "how-to" aspect is what uniquely differentiates design theory from other types of

theory: "[Design theories] give explicit prescriptions on how to design and develop an artifact, whether it is a technological product or a managerial intervention" (Gregor and Jones, 2007, p. 313).

Cook and Campbell (1979, p. 28) articulated the need for design theories ("recipes") in the social sciences:

> Knowledge of causal manipulanda, even the tentative, partial and probabilistic knowledge of which we are capable, can help improve social life. What policy makers and individual citizens seem to want from science are recipes that can be followed and that usually lead to desired positive effects, even if understanding of the micromediational processes is only partial and the positive effects are not invariably brought about.

Due to its pragmatic appeal, researchers condemned design research to be unscientific and easy to confuse with consulting work (Kasanen and Lukka 1993). However, as Gregor (2002, p. 18) argues based on the design theory presented by Markus et al. (2002):

> [Design theory] consists of general principles to solve a class of business problems, rather than a unique set of system features to solve a unique business problem. [This] abstraction and generalization […] distinguishes [design research] from what would occur in consulting.

A further distinction of design research from consulting is the use of kernel theories (Walls et al. 1992) and justificatory knowledge (Gregor and Jones 2007). Kernel theories are reference theories that inform the design of artifacts or help to explain and justify their practical utility. They often come from disciplines outside of IS, such as the natural sciences or social sciences. Kernel theories are a form of justificatory knowledge, but they are broader, including also the experienced-based tacit knowledge of practitioners (Gregor and Jones 2007).

According to Merton (1968), design theories are mid-range theories that

"lie between the minor but necessary working hypotheses that evolve in abundance during data-to-day research and the all-inclusive systematic efforts to develop a unified theory that will

explain all the observed uniformities of social behavior, social organization, and social change [, grand theory]."

Merton (1968) and Cook and Campbell (1979) emphasize that applied fields like IS or IT should focus on such mid-range theories.

The concept of artifacts in design science can appear fuzzy. This may be due to the fact that some types of design theory, e.g., development methods, can also be regarded as artifacts (Gregor and Hevner 2013). Goldkuhl (2002) proposes a rather strict definition of artifacts: "an artefact […] is used to perform material acts by virtue of its material properties." (p. 4) Gregor and Hevner (2013), on the other hand, give a loser definition: "the term *artifact* is used […] to refer to a thing that has, or can be transformed into a material existence as an artificially made object (e.g., model instantiation) or process (e.g., method, software)" (p. 341).

Following this definition, a theory is different from an artifact in so far as it embodies knowledge that goes beyond "the description of a materially existing artifact." (Gregor and Hevner 2013, p. 341)

Gregor and Hevner (2013) discuss three different abstraction levels of design science knowledge contributions:

(1) situated implementations of artifacts (no theory),
(2) nascent design theory (components of theory), and
(3) well-developed design theory about embedded phenomena (full-blown theory; see also Gregor and Jones, 2007).

Gregor et al. (2020) argue that design principles in the form of prescriptive statements are the necessary, and most distinctive, but not the sufficient condition for design theory. Different views on the constituting elements of design theory exist, of which Walls et al. (1992) and Gregor and Jones (2007) are examples. While they are different in some points, they have some commonalities. Both of them state that a design theory should present at least:

(1) A description of the goals and the class of systems that the design theory aims at

(2) Design Principles in some form

(3) Falsifiable propositional statements

(4) Justificatory knowledge or kernel theory

(5) The description of an implementation method

Furthermore, Gregor and Jones (2007) specify: "[a]s the word 'design' is both a noun and a verb, a theory can be about both the principles underlying the form of the design and also about the act of implementing the design in the real world (an intervention)" (p. 322).

4.4 Qualitative Research

To achieve research goal 4, we jointly started a study of AI value creation mechanisms along the data-to-insight-to-action-to-value path (see Sharma et al. 2014). For this study, we followed an exploratory design based on semi-structured interviews and analyzed the data qualitatively. We conducted two rounds of data collection. During the first round from October 2018 to January 2020, we interviewed 40 data scientists to get their retrospective accounts and bottom-up inside views (Plastino and Purdy 2018) from designing and implementing 57 organizational applications of AI, while they were specialized differently, their main specialization was implementing machine learning algorithms. During this study, we gave our "informants" an "extraordinary voice" (Gioia et al. 2013, p. 26) as we treated them like "knowledgeable agents" that "know what they are trying to do and can explain their thoughts, intentions, and actions" (Gioia et al. 2013, p. 17). During the interviews, we started by asking our informants how they approached value creation with their applications of AI and in particular, how they approach turning from questions to the data or from data to questions, form data to knowledge, from knowledge to decisions, and from decisions to actions (Sharma et al. 2014; Thiess and Müller 2020a). However, with the help of a logbook, we adjusted our interview guide based on our theorizing and the informant's accounts (Gioia et al. 2013). Here we focused on their perception of value contributing mechanisms, enablers, and value diminishing challenges. The interviews are, on average, 50 minutes long. We transcribed the recorded interviews in more than 600 pages.

Between January and June 2020, we conducted a second data collection round that focused on the validation of our findings. In particular, we held a focus-group workshop with 13 informants (seven earlier informants and six new). During the workshop, we presented our findings and asked them to discuss and reflect upon them based on their field experiences. Moreover, we got substantial written feedback on the findings from ten informants.

During the study, we conducted two different overall analyses that resulted in two different papers (*Paper 1* and *Paper 2*). In the first analysis, we stayed closer to the original interview guide and, based on this, conceptualized an AI value creation process along with seven critical challenges in each sub-stage of the process. In the second analysis, we re-adjusted the initial research guide more vividly, which resulted in the conceptualization of AI value creation mechanisms, corresponding AI system types, and necessary but not sufficient conditions for implementing them. Moreover, during the analysis, we focused on how firms shift between the different value creation mechanism in response to changed objectives but mainly due to matches and mismatches in the conditions

Our data analysis followed Gioia et al. (2013) as we analyzed the data in several rounds, of which the first resulted in a large number of open codes that are close to the original in vivo formulations of the informants. In later rounds, we used our experience and theoretical knowledge to condense our emerging concepts into themes and dimensions. Here, we tried to let the themes and dimensions emerge from the open codes and not to force them into pre-existing frames. Figure 12 shows an example of our coding for the value creation mechanism analysis.

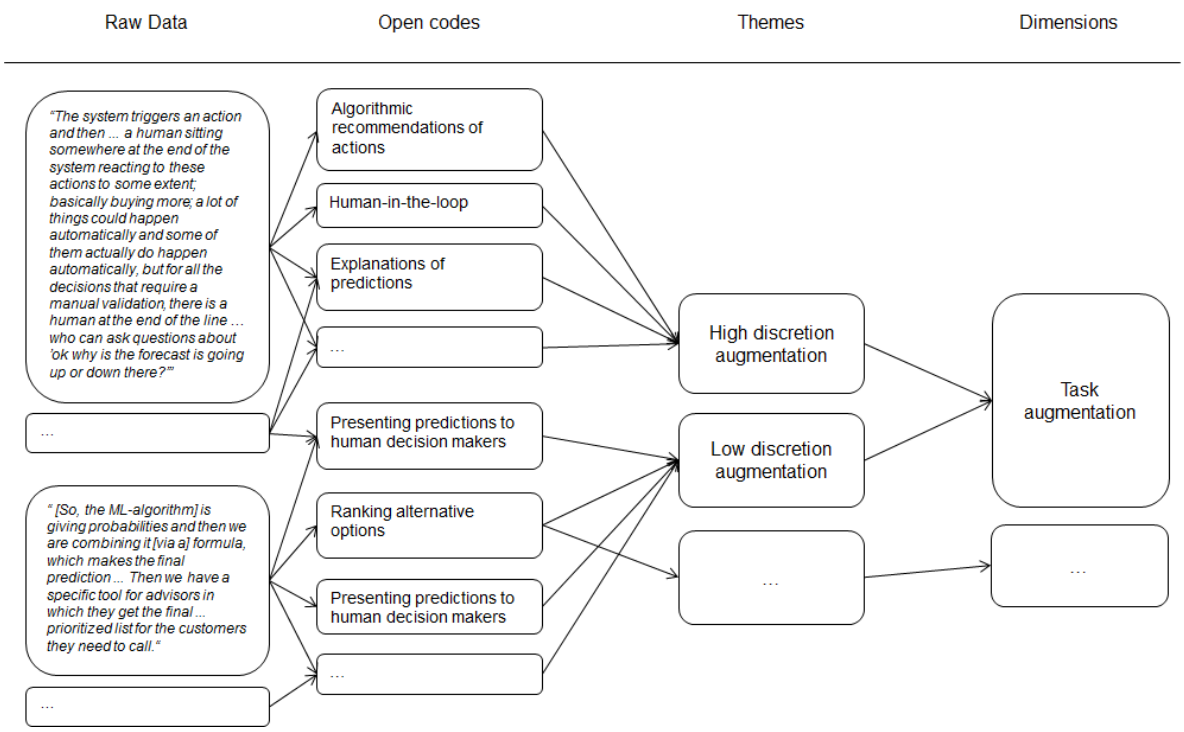

Figure 12. From raw data over open codes to themes and dimensions

5 Summary of Results

The chapter summarizes the findings from the different studies that this PhD thesis encompasses. Figure 13 illustrates how the behavioral AI value creation study informed the ADR program. We identified the need for this behavioral study during the literature search for ADR Study 1. For this ADR study, the AI value creation process mostly informed the further development of a method for data-driven lead generation (based on AI System 1) that we published in Paper 4. For ADR Study 2, the behavioral research results informed the search for justificatory knowledge about its interpretability features and design principles. For ADR Study 3, the AI value creation study informed the problem formulation, the initial solution design, and the search for justificatory knowledge.

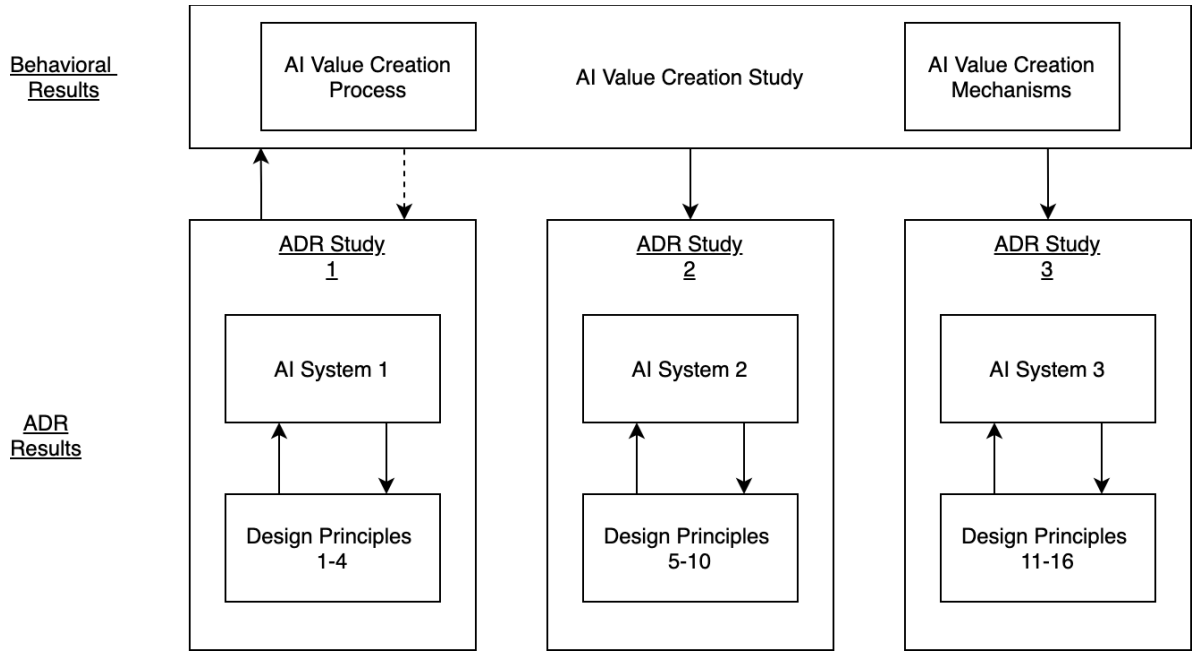

Figure 13. Overview of research results

5.1 AI Value Creation Study

This study resulted in two papers, Paper I and Paper II. Paper I presents a conceptualization of the AI value creation process along with critical challenges in its sub-stages. Moreover, it suggests a selection of enablers to address the challenges. Paper II presents a conceptualization of AI value creation mechanisms along with a description of how firms shift through such mechanisms due to the match or mismatch of necessary but not sufficient conditions for realizing their value propositions.

5.1.1 Challenges Along the AI Value Creation Process

Figure 14 shows a conceptualization of the AI value creation process. The process consists of phases for planning, developing, and operating AI processes. An AI process is a decision making process that an AI system supports or fully executes. It involves sensing or collecting data, turning data to knowledge with the help of machine learning algorithms, turning knowledge to decisions by representing the knowledge in a user-friendly way via user interfaces, or by automatically making decisions with the help of artificial choice models (utility functions or optimi-

zations). The Decisions to Actions phase is about implementing decisions via physical interventions into the environment performed by human or machine actors.

The Planning phase is concerned with problem definition, the initial design of AI solutions, project management activities, and overall project governance. The planning phase reflects the need for dynamic readjustments to initial project plans due to the high outcome uncertainty involved in developing and implementing advanced AI solutions. Therefore, it is continuous and runs in parallel to development, AI process execution, and operations. The development phase is closely connected to the planning phase and involves activities such as constructing databases, data pipelines, machine learning models, user interfaces, or interfaces to transactional systems. Moreover, it often involves the development of prototypes used to seek management buy-in for proceeding to integrate an AI system with an organization's IT architecture. The operations phase is about continuously monitoring the AI process' performance and, based on that, to improve it or, when necessary, to repair it.

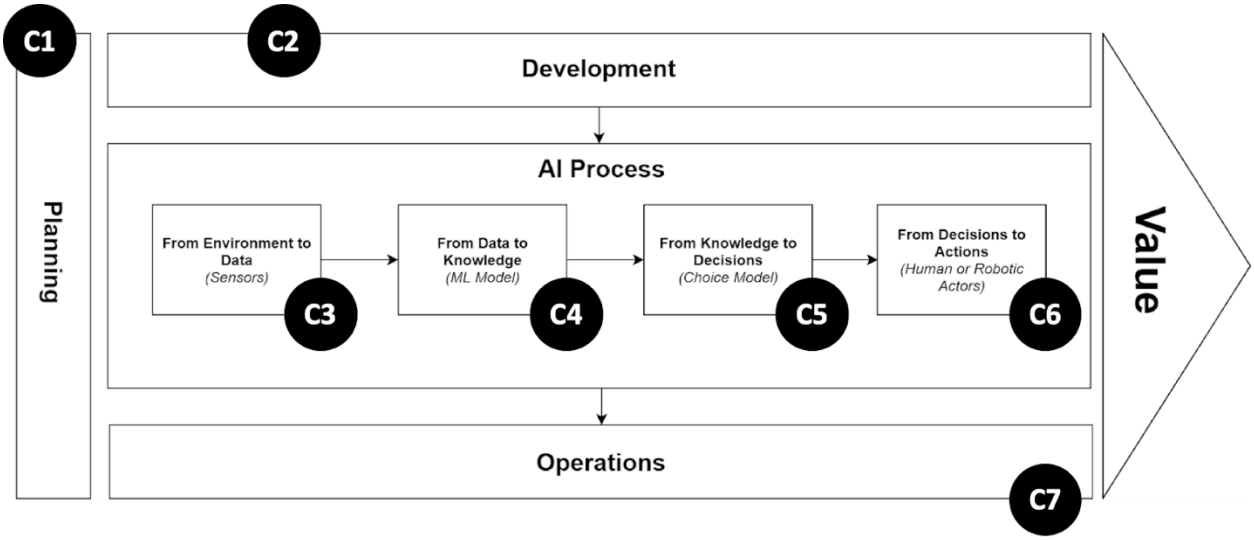

Figure 14. Challenges along the AI value creation process

As indicated in Figure 14 with a capital C, we identified critical challenges for each phase of the Al value creation process. C1 is called Inflated Management Expectations and about managers that assume that one can automate all kinds of knowledge with Al similarly to how one can automate less knowledge-intensive work with technology. Another instance of this challenge is managers that "think big" and want to go for "moonshot" projects right away. Instead, in the

planning phase, one should rather assess whether a task is suitable for Al; in some cases, engage a consultancy or an analytics vendor, or conduct Al pilot projects first.

C2 refers to Managing Al Projects like Traditional IT Projects. It is about managers that neglect the inherent outcome uncertainty of Al projects, and, due to this, expect data scientists to deliver results and "quick-wins" continuously. Another instance of this is that data scientists feel stressed and pressured to draw conclusions early and report intermediate findings without feeling sure about them. To avoid this, one needs managers with good AI skills that are able to realistically assess the potentials of AI projects, set appropriate KPIs, and understand how much effort it takes data scientists to create credible insights and stable AI systems. Moreover, data scientists can help to make managers more AI savvy by communicating their AI related work processes well.

C3 refers to Data Availability and Quality. This challenge is about the lack of appropriate training data. Even if companies nowadays have access to big data, it does not equate high-quality data. Moreover, even after often spending more than 80% of their work time on data processing and feature engineering, it can be hard to find the ”signal” in the ”noise.” A remedy could be to collect new, high-quality data in the field, e.g., via sensors, or by using Amazon Mechanical Turk for data labeling. Also, a company-wide data culture helps to assure a higher quality, e.g., by making users of transactional systems aware that the data they enter is an asset that can be fortified via AI systems and processes.

C4 (Interpretability) refers to the problems for user understanding, acceptance, and learning that are caused by the black-boxed ways in which many machine learning algorithms train models. In the case of the Random Forest algorithm (Breiman 2001a), for instance, one cannot access the rules based on which a model makes predictions. While this is possible for traditional decision tree-based models, Random Forest is based on an ensemble approach in which the algorithm averages the predictions of many slightly different decision trees. This lack of interpretability creates a situation in which only the AI system learns but not its human developers and end-users. Potential solutions to this problem are intrinsically interpretable algorithms that

fit a function of beta coefficients that enable humans to even predict with them with pen and paper, such as linear and logistic regression, but also decision trees. The downside of these models is, however, that they are less accurate. Due to this, researchers started to develop post-hoc explainability methods (Scott M. Lundberg et al. 2018; Lundberg and S.-I. Lee 2017; Lundberg and S. I. Lee 2017; Ribeiro et al. 2016a, 2016b) that basically use the trained machine learning model for input-output simulations to estimate its underlying rules.

C5, Causality, refers to the challenge that is caused by the fact that people tend to confuse that association does not equal causation. Associational analyses often suffer from spurious correlations that are resulting from confounding variables that both affect the features, and the target variable of a machine learning model (see Pearl and Mackenzie 2018). To approach this challenge, randomized controlled trials (Cochrane 1972), and other quasi-experimental methods to causal inference (Cook et al. 1979; Varian 2016) can cope with many of such issues. However, such analyses have the aim to ”explain” rather than ”predict” (Shmueli 2010). Nevertheless, causal modeling techniques that use diagrams and data scientists’ subject-matter knowledge to uncover confounders and potentially control for or condition on them can improve the validity of predictive models too.

C6, Missing Link to Transactional Systems and Processes, refers to situations in which it is, for technical, organizational, or legal reasons, hard or not possible at all to feed the knowledge that AI systems create back into the transactional systems on which most of the organizational processes rely. As a result of this difficulty in accessing the AI-system-generated-knowledge, employees may simply ignore them. In such situations, robotic process automation can help to connect unconnected systems manually. However, they need to be developed, maintained, and operated too. Also, optimization algorithms that output not only knowledge, e.g., by scoring a set of alternatives, but make decisions (choosing the best alternative), can potentially “[make] it harder for decision makers to avoid using analytics - which is usually a good thing” (Davenport 2013).

C7, Dynamic Environments, refers to the challenge that arises when the performance of AI systems diminishes, due to, e.g., unnoticed changes in the input data, or cases in which the system was implemented by a consultancy, however, once they implemented the AI system, there were no employees that could adjust or retrain the model. A potential way to deal in particular with input data related changes are machine learning operations that can entail monitoring the performance of AI systems continuously so that one can intervene and repair or improve the system when one observes a drastic drop in predictive performance.

5.1.2 Shifting AI Value Creation Mechanisms

As the second main result of this behavioral study, we identified three value creation mechanisms. To create value through AI Value Creation Mechanism 1, Knowledge Creation, organizations develop and implement knowledge creation AI systems that one often uses in data-intensive research projects. Such systems deploy machine learning algorithms to create new knowledge in the form of patterns and rules. The value target of such systems is organizational knowing (Shollo and Galliers 2016). Common types of machine learning that these systems use are hypothesis testing, unsupervised clustering approaches, simulation, and causal inference. Usually, knowledge creation systems support non-programmable decisions on a tactical and strategic level. For such systems, both the decision-maker and action taker are typically humans.

One approaches AI Value Mechanism 2 with task augmentation systems, such systems train supervised machine learning models via a variety of algorithms such as random forest (Breiman 2001a) or neural networks (Bishop 1995). A task augmentation system makes predictions on unseen data based on the rules and patterns that the trained model learned on old data. The predictions generated by these systems support human decision-makers in executing a programmed task on a tactical and operational level. We differentiate two task augmentation subtypes, low- and high discretion systems. The outputs of a high-discretion task augmentation system leave some room for human influence on the eventual decision about what particular course of action one should take. Low-discretion augmentation systems, however, commonly use a decision function that already makes the decision about what alternative course of action

one should take, leaving human decision-makers only with the choice to accept or reject the decision that the system made. Also, in some cases, humans are in the loop only to control and make sure that the system works error-free, in situations where the trust in such systems is low, or one has to meet certain legal requirements.

AI Value Mechanism 3, Process Automation, one achieves with AI process automation systems. Their value targets are increased productivity through process exploitation, but also creating new value offerings through intelligent products and services, e.g., smart chatbots. The main outputs of such systems are prescriptions as they utilize machine learning models to make predictions and scores about alternative decision options and use decision functions to find the best option. However, in contrast to the other system types, they do not support human decision making tasks directly. Instead, the prescriptions that AI systems generate instruct via interfaces machine action takers in executing operational processes fully automatically. See Figure 15 for an overview of the AI system types.

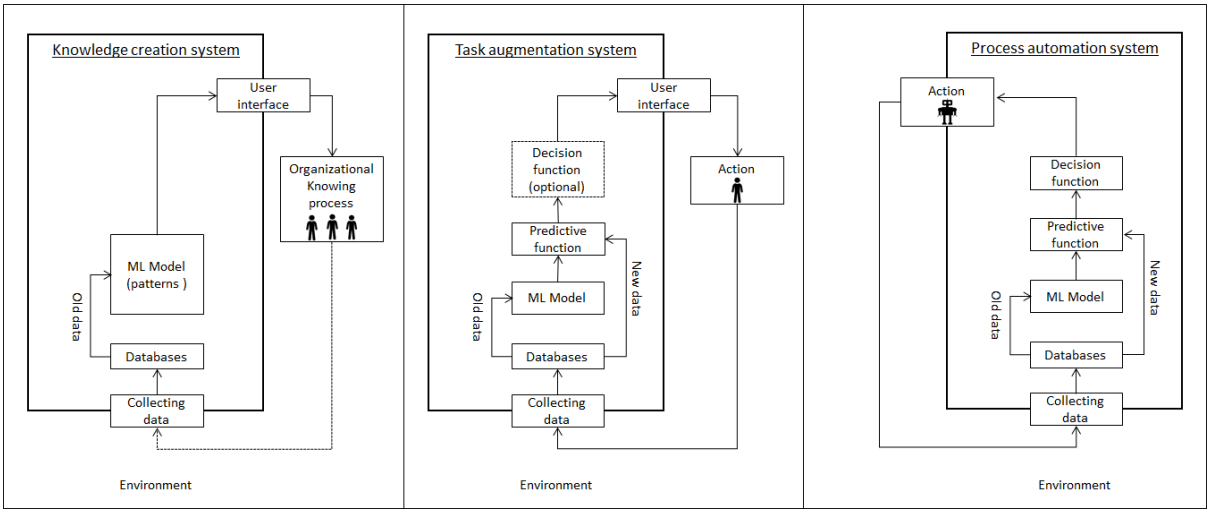

Figure 15. Differences between knowledge creation, task augmentation, and process automation systems

Moreover, we found that organizations shift between AI value creation mechanisms. Such shifts, we explain with a match or mismatch of necessary but not sufficient conditions for the configurations of each value creation mechanism (see Figure 16). In one application of AI, for instance, a large jewelry retailer had already successfully implemented a process automation

system for their online advertising. However, due to the outbreak of the Covid-19 pandemic, the assumptions, and rules that the underlying model was based upon did not explain customer behavior well-enough anymore to create accurate predictions. This unstable environment required them to re-assess the situation, and they started again with a data-intensive research project that was based on data from countries that the pandemic hit first. Based on the increased understanding of the situation and learnings from the data-intensive research project, they trained a new predictive model that supported a human decision making task-force to bring their online marketing back on an acceptable level. Eventually, when the environment is more stable again, they want to turn the high-discretion task augmentation system into a process automation system again. Figure 17 illustrates this reconfiguration of AI value creation mechanisms (project 56).

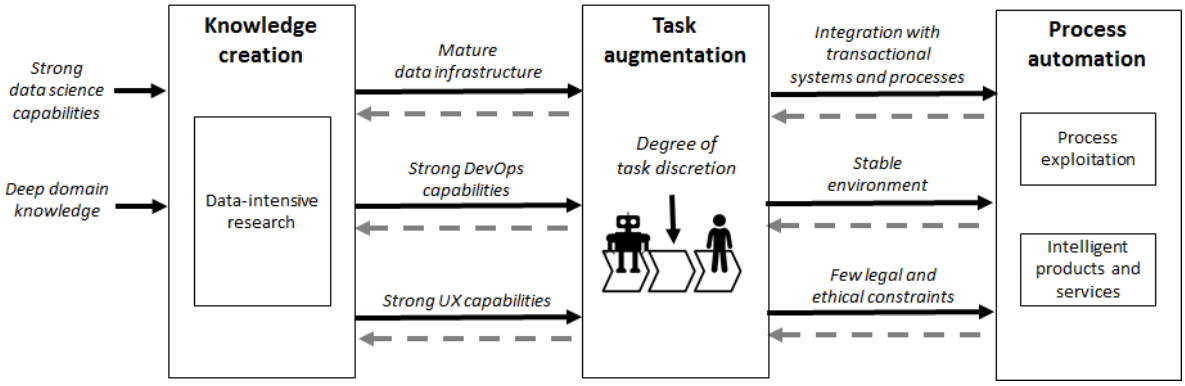

Figure 16. AI value creation mechanisms and their necessary but not sufficient conditions

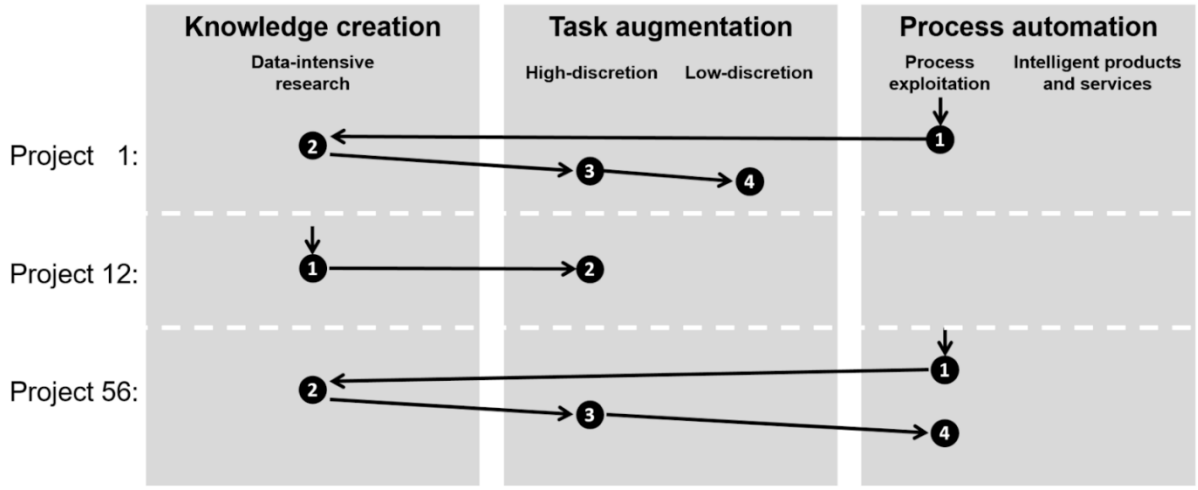

Figure 17. Selected reconfigurations

5.2 ADR Study 1: Data-driven Lead Generation

5.2.1 Problem Formulation

In reaction to an increasingly challenging market, MAN Energy Solutions had recently implemented a new customer relationship management system to support the transformation towards a customer-centric and pro-active aftersales approach. However, we diagnosed that the system was under-used due to a lack of pro-active work practices. We identified this field problem as a research opportunity to investigate:

How can AI systems enable proactive customer relationship management processes?[2]

We structured our problem formulation and solution design with the data-to-insight-to-value conceptualization by Sharma et al. (2014). Especially, we searched the literature for common challenges in data-driven value creation. Moreover, the ML model development was guided by the cross-industry standard process for data mining (CRISP-DM; Shearer et al. 2000). Also, the information-decision-insights-supervision framework by Dearden (2001) guided the partially automated features of our solution artifact.

[2] In the paper, we talked about data-driven decision making as defined in Section 2.1

Throughout the cycles of building, intervention, and evaluation, the problem formulation, but especially the latent solution design, was shaped and informed by different theories that helped us explain and/or solve unanticipated problems and solutions. The theory of Occam's Razor that says: "Given two explanations of the data, all other things being equal, the simpler explanation is preferable" (Haussler and Warmuth 1987), informed our design decision of moving from a highly complex and uncertain data-driven machine learning approach to a well-established approach to machine learning that is grounded in marketing theory (see Platzer and Reutterer 2016 and DP 1-3). Also, the No-free Lunch Theorems (Wolpert and Macready 1996), which imply that no one algorithm is best for all problems, helped us to explain why in spite of using a generally well-performing algorithm such as random forest (Breiman 2001a), we could not reach a satisfying performance.

Moreover, we used arguments by Watson (2014) and LaValle et al. (2011), who argue that firms should use analytics to prescribe action, to justify our idea of formulating the descriptions of leads prescriptively when presented to sales professionals (see DP4).

Also, we used the 3-Gap Framework by Kayande et al. (2009) to explain how complexity reduction, e.g., when following Occam's razor, affects comprehensibility, and acceptance (see DP2-3), and we used Hollander et al. (1973) to explain and justify the effects that we observed when incorporating domain knowledge into the system.

5.2.2 Building, Intervention, and Evaluation of Solution Artifacts

The concrete design goal was:

To construct a system for generating personalized and data-driven aftersales service leads by utilizing AI technology and customer-life-cycle data

In reaction to this, throughout many BIE cycles, we developed the Data-driven Lead Generation method and a situated implementation of the respective AI system. We formalized its third main iteration (see Figure 18 and Thiess and Müller 2018) based on a concrete implementation of a data-driven lead generation system for shipowner changes at MAN Energy Solutions. The system utilizes internal transactional data as well as external databases of ship metadata. The

system was executed monthly. Here, a lookup algorithm compared the shipowner data column of the most current version of an external database with a saved version of the database from the month before. If it detected deviations in the shipowner data column, it assumed that the owner of a ship had changed. Such an owner change constituted a lead event. To be able to prioritize such lead events, we calculated customer lifetime values with the help of hierarchical Bayesian probability models (Fader and Hardie 2009; Platzer and Reutterer 2016) for each customer (shipowner). This allowed sales professionals to select and reach-out only to the customers with the highest future customer lifetime values as reaching-out to a customer that has a very low customer lifetime value, and has, thus, probably churned, can be a waste of scarce resources. Moreover, the system created additional descriptive reports of relevant sales and metadata that we attached to the lead object in the customer relationship management system were the system assigned it to a sales professional.

In our second publication about this ADR study (Thiess and Müller 2020a), informed by the AI Value Creation Study (see Section 5.1), we abstracted the Data-driven Lead Generation method further and changed its representation by, e.g., integrating it with parts of the AI Value Creation process and adding guiding questions and into each stage (see Figure 19).

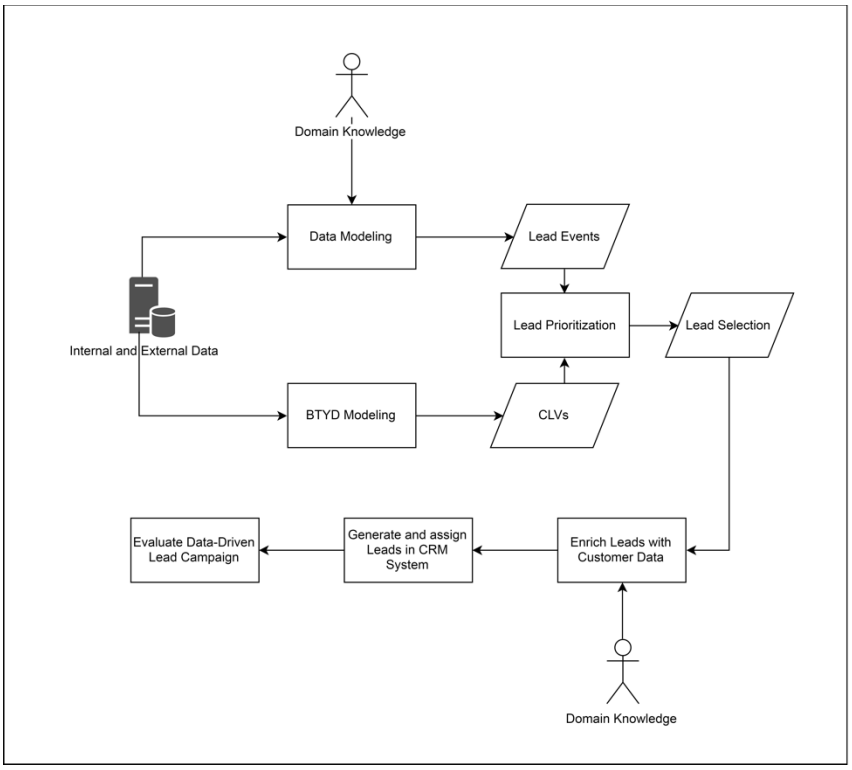

Figure 18. Data-driven lead-generation artifact after the third iteration

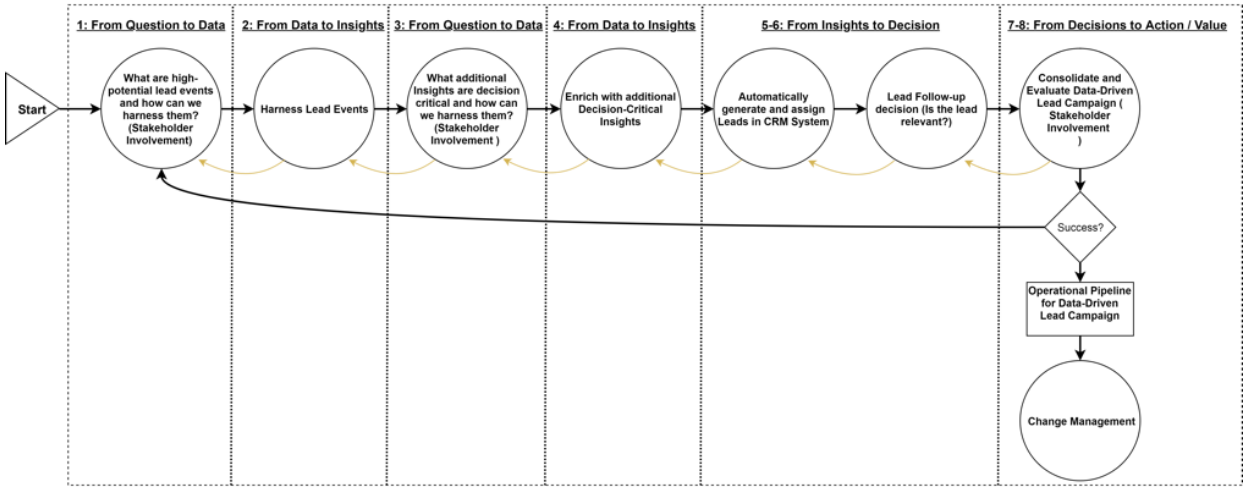

Figure 19. The final data-driven lead generation artifact

In our instantiation of the data-driven lead generation method, we calculated customer lifetime values via so-called buy-till-you-die models that belong to the class of hierarchical Bayesian probabilistic models (Fader and Hardie 2009; Platzer and Reutterer 2016). From this model family, we eventually selected the Pareto/GGG model by Platzer and Reuterer (2016), which showed the lowest error rates and allowed us to incorporate purchasing regularity parameters into the model. The model makes predictions by simulating posterior probability distributions of being active and alive in the future for each customer. Here, we used an additional Gamma-Gamma model, as described by Fader and Hardie (2013).

During the many BIE cycles, we evaluated the artifacts continuously, e.g., based on structural, functional, and usability related factors. The main criterion, though, was effectiveness (did the artifact achieve its goals?). We evaluated the method mostly by demonstrating the effectiveness of the implemented data-driven lead generation system at MAN Energy Solutions. We could demonstrate that the systems created data-driven leads that sales professionals used at MAN Energy Solutions. In particular, they used the leads to approach high potential customers with personalized offers based on their interaction history. When we started the project, sales and marketing were mostly reactive (e.g., a customer requested overhaul and spare parts services after an engine breakdown) and with the data-driven lead generation, we could effectively demonstrate how to utilize the new customer relationship system better and to create pro-active and customer-centric aftersales service processes.

5.2.3 Reflection, Learning, and Formalization of Design Principles

We developed a set of design principles that we abstracted to the broader class of systems for data-driven decision making. However, this broader class contains the narrower class of AI systems for B2B aftersales decision support.

DP1: Theory-driven modeling – Given a lack of proof-of-concept, use theory-based models instead of data-driven machine learning algorithms to achieve concrete results.

We designed the initial iteration of the data-driven lead generation method around data-driven machine learning algorithms like gradient boosted trees (Breiman 1997; Friedman 2001, 2002).

At this point, the goal was to create leads by predicting when the next major overhaul of a ship engine was due. However, we soon realized that the approach was complex, and the data was not sufficient (see Haussler and Warmuth 1987; Wolpert and Macready 1996), and most importantly, we lacked theoretical guidance as no one else had ever made this approach either in theory or practice. We realized that the project had a high risk of failure due to its high outcome uncertainty. So eventually, we decided to discontinue the approach, leaving us without a concrete solution for the field problem.

In the following, we looked into the scientific marketing literature to find suitable, and well-described approaches for creating and or prioritizing sales leads at MAN Energy Solutions. Here, we selected a probabilistic hierarchical Bayes approach to predict customers' future purchasing behavior and to calculate customer lifetime values (Fader and Hardie 2009; Platzer and Reutterer 2016). The method was based on well- theorized and proven parametric assumptions about the model input and outputs, and, thereby, allowed us to reduce the outcome uncertainty. Finally, we designed and instantiated the data-driven lead generation method around this approach. We concluded that when there is high outcome uncertainty, due to the lack of proof-of-concept, data scientists should choose established theory-based models to achieve concrete results with high certainty.

DP2: Comprehensibility – Limit models' complexity to gain support from managers.

According to Gregor and Benbasat (1999), the comprehensibility of decision support systems plays a major role in technology acceptance. Our initial black-boxed machine learning-based lead generation approach was not well-received by managers and other relevant stakeholders. We observed that they struggled to understand the inner working of the model. This was due to the high complexity involved, e.g., using a large amount of data with many different variables as input for a black-boxed ML algorithm in an approach for which no theoretical or practical guidance existed.

In contrast to that, the eventually implemented BTYD models for calculating CLVs (Fader and Hardie 2009; Platzer and Reutterer 2016) were much better perceived. While they are still complex due to the Bayesian estimation procedure, we observed that stakeholders were much more accepting of them. We explain this by their simple overall structure that requires only three factors to calculate CLVs: the time of a customer's last transaction (recency), the total number of a customer's purchases (frequency), and the monetary value of purchases.

DP3: Domain Knowledge – Incorporate domain knowledge into the data-driven decision-making process to encourage acceptance by managers.

We incorporated different forms of domain knowledge into the data-driven decision making process that the instantiation of the data-driven lead generation method enables. First, we reached out to domain experts to investigate what events surrounding the customer and ship engine-lifecycle in MAN Energy Solution's aftersales could determine a potential lead. Then we operationalized this knowledge in the form of explicit business rules like "If the owner of a ship changes, generate a lead and assign it to the responsible sales professional." Moreover, the buy-til-you-die modeling approach that we choose to calculate customer lifetime values allowed us to incorporate informative prior parameters that one can adjust based on domain knowledge. We justify our approach by referring to Hollander et al. (1973). They argue that letting stakeholders participate in problem-solving (in our case by inscribing their knowledge into an IT artifact), should increase their acceptance towards the approach.

DP4: Actionability – Provide actionable insights instead of quantitative reports to increase use by decision-makers.

LaValle et al. (2011) show that many applications of AI systems lack prescriptions for concrete actions to their users. Practitioners of the business intelligence department at MAN Energy Solutions were mentioning early on that those applications that supported or improved an already existing business process were much more frequently used than those for which they tried to build a new business process around it. Based on this insight, we decided to display data-driven leads into the newsfeed of the customer relationship management system instead of building a new interface. Then, if sales professionals clicked on the lead, they got a clear pre-

scription of what to do, like contacting a particular customer because the ownership structure of one of its vessels changed. In addition to that, we attached reports of key customer and vessel metrics directly to the lead object in the customer relationship management system. This way, sales professionals had all the necessary information at their fingertips and could take action on the prescribed lead immediately. Thereby, the system "makes it harder for decision-makers to avoid using analytics—which is usually a good thing" (Davenport 2013).

5.3 ADR study 2: Sales Win-propensity Prediction

5.3.1 Problem Formulation

MAN Energy Solution wanted to increase its conversion rate of aftersales service quotations into sales orders. In aftersales, one usually first receives a request for a quotation based on which the OEM prepares a sales quotation by checking material availability, prices, and potential discounts. As it is not unusual that customers request a sales quotation from several competing OEMs, one needs to actively follow-up and engage with the customers to improve the win-propensity. This follow-up process is, however, resource-intensive, and in some cases, the probability of converting such a sales quotation into a sales order is low.

We identified this situation as a research opportunity to investigate:

How to design AI systems that support resource allocation decisions in aftersales quotation follow-up and help to increase the conversion rate?

The general body of knowledge about supervised machine learning informed our initial solution design and, in particular, theory about logistic regression (e.g., Friedman et al. 2001) and ensemble decision tree-based approaches (Breiman 2001a). Moreover, the AI Value Creation study structured and guided our theory search.

Throughout the cycles of building, intervention, and evaluation, the problem formulation, but especially the latent solution design, was shaped and informed by different theories that helped us explain and/or solve unanticipated problems and solutions. Theories of AI acceptance and explainability (e.g., Gregor and Benbasat 1999; Kayande et al. 2009; Lundberg and S. I. Lee 2017;

Martens and Provost 2014) helped us to design and explain local-contrastive features (DP1) of the AI system. Based on such theories in combination with theories on accountability (e.g., Lerner and Tetlock 1999) and causality (e.g., Wheelwright et al. 1998), we built related features for global explainability, accountability, and causality (DP7-8 and DP10) into the AI system. Moreover, theory on data representation (e.g., Larkin and Simon 1987; Miller 1956) guided and justified us in displaying only a selection of model explanations (DP6), while cognitive decision making theory informed our knowledge representation more generally (e.g., Thaler and Sunstein 2009).

5.3.2 Building, Intervention, and Evaluation of Solution Artifacts

The concrete design goal was to construct a system that can predict the win-propensity of an open aftersales quotation based on transactional and customer data. Such an approach would allow firms to focus resource allocations in personalized quotation-follow up on high-potential customers.

After many BIE cycles, we designed and implemented a system that is based on a highly efficient gradient boosting machines algorithm called lightGBM (Ke et al. 2017), a second-level model of quotation age, and SHapley Additive exPlanations (Scott M Lundberg et al. 2018; Lundberg and S. I. Lee 2017; see Figure 20).

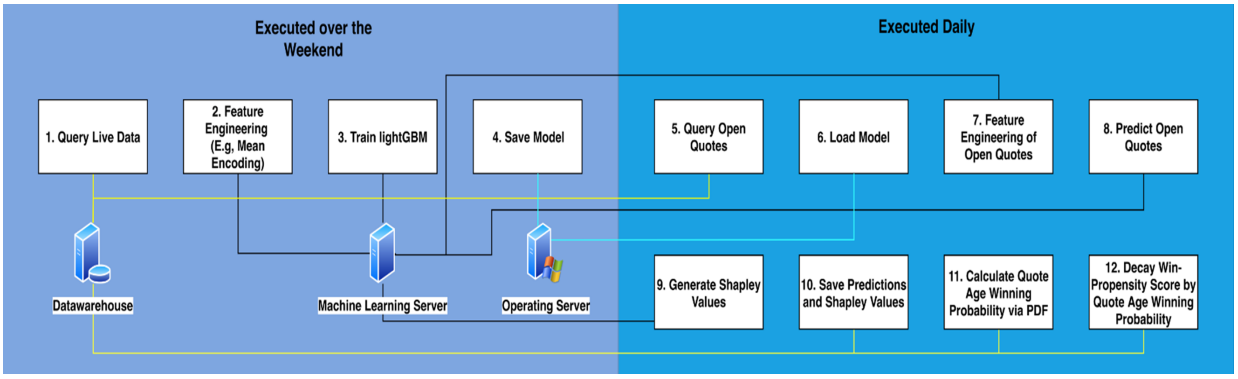

Figure 20. Implemented back-end process

To address the challenge that the win-propensity of an open quotation decreases over time, we incorporated a second level probability model of quotation age. Here, we use a probability den-

sity function that represents the historical distribution of converted quotes with the percent of converted quotes on the y-axis and the number of days from open quote to conversion on the x-axis. We then decay the raw win-propensity predictions by the second-level probabilities. Then, we display both scores in the user interface.

In addition, we implemented a version of the SHapley Additive exPlanations algorithm (Scott M. Lundberg et al. 2018; Lundberg and S.-I. Lee 2017; Lundberg and S. I. Lee 2017) that utilizes a concept stemming from game-theory in which one attempts to determine the marginal contribution of one player to the success of the whole team. This approach was reused to determine the marginal contribution of one variable to the overall prediction. The SHapley Additive exPlanations algorithm enables local-level explanations (Shapley values), which means that one can display for each prediction, the top contributing variables (see Figure 21). Moreover, one can take the average of local Shapley values to gain more global insights into how the model works, and what variables drive the sales conversion process.

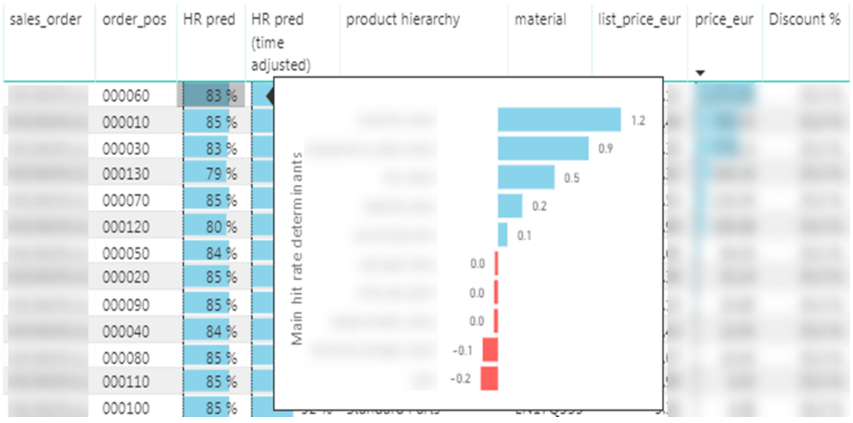

Figure 21. Local instance-level SHAP explanation (blurred for confidentiality reasons)

Next to the artifacts, we formalized an implementation method (see Figure 22) by reflecting on the BIE cycles during this ADR project. The method guides others on a general level to implement similar systems in organizational contexts.

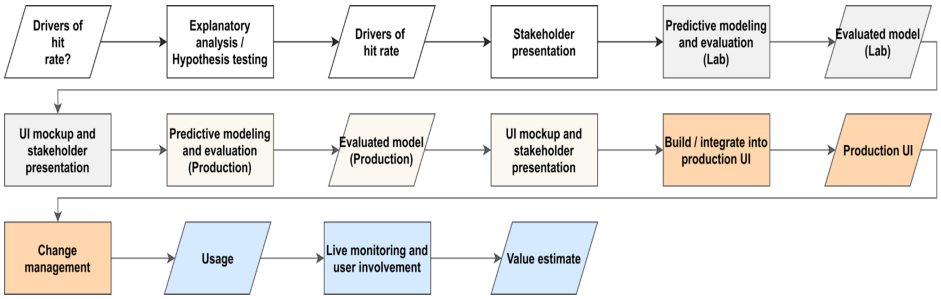

Figure 22. Implementation method

We evaluated the artifact continuously in terms of structural, functional, and usability related features. However, the main evaluation criterion was effectiveness. The feedback of key stakeholders at MAN Energy Solutions was very positive. The system was amongst others labeled by one of the managers as the best approach to AI aftersales service support that she had seen so far. Also, the end-users accepted the system outputs well, and currently, around 50 active users work with it daily. Sales professionals and managers indeed liked it so much that it is now planned to implement a similar system not only for the Danish headquarter of MAN Energy Solutions, but for the German headquarter as well.

5.3.3 Reflection, Learning, and Formalization of Design Principles

We developed a set of design principles that we abstracted to the class of sales win-propensity prediction systems. They may, however, apply even to the broader classes of AI systems for B2B aftersales decision support and explainable AI systems in general.

DP5: Local-contrastive Explainability – Present model explanations to users on an instance-level to support contrastive explanation processes

Sales professionals at MAN Energy Solutions utilize their deep domain knowledge to address challenges such as large portfolios of heterogeneous products or intransparent owner structures of ships when doing their work. We observed that some of the sales professionals did not trust the predictions of machine learning models. Here, they are less interested in a detailed description of how a model generated a prediction, but rather in why a model made a certain prediction (e.g., 90% win-propensity) and not another (e.g., 20% win-propensity). Lipton (1990) and Miller (2019) describe such explanation processes as "contrastive." They argue that people com-

pare a particular event (or prediction) with an imagined counterfactual event to make sense of it. Harman (1965), argues for an explanation as an abductive process, in which people make sense of observations by "inference to the best explanation." In our user interface, we display instance-level Shapley values to help users find a satisfying explanation for why score X and not score Y by allowing them to compare their own domain, knowledge-based mental models, with the inner workings of the machine learning models (see Kayande et al. 2009).

DP6: Selective Visualization – For local explanations, visualize only the top contributing features to reduce explanation complexity

We display the win-propensity scores in a general quotation follow-up tool that contains already a lot of information. To not overload users, and reduce the complexity of explanatory Shapley values, we only show the top features that contribute to a particular prediction. We justify our empirically motivated design principles by referring to research from the field of psychology that suggests presenting information in 4-7 chunks, which is how they argue, the largest information unit that humans can retain in short-term memory to process it (Cowan 2001; Larkin and Simon 1987; Miller 1956).

DP7: Accountability – Schedule regular management presentations to increase data scientists´ need for justification

We experienced that scheduling regular meetings with management as the key stakeholders in the system approval process helps to keep them involved and thereby to anchor the project in the organization. However, we also realized that such meetings prompted us as developers to work in a more structured and transparent way. Research suggests that having to justify one's views and actions can have debiasing effects and lead to work practices that are more evidence-based (see Simonson et al. 1992; Tetlock 1985; Tetlock et al. 1989a). However, certain kinds of accountability, like outcome accountability, can have detrimental effects too (Lerner and Tetlock 1999). Therefore, one should instantiate this design principle carefully.

DP8: Global Explainability – Explain the machine learning model to managers on a global level to increase acceptance, enable process accountability, and share outcome accountability

We observed that managers became much more engaged and positive towards our system ones they saw the results of our explanatory analysis of the main drivers of conversion rate. Also, the display of such global explanations (average Shapley values) allowed them to understand better how the model works and what our process for developing the system was. This increased understanding allows managers to evaluate the performance of the system and us as its developers based on the underlying processes rather than on the outcomes of the model only. Such a form of process accountability can alleviate some of the negative effects that outcome accountability can create, and lead to a more thorough and transparent decision making (Lerner and Tetlock 1999; Simonson et al. 1992). At the same time, when managers understand the process of developing and implementing an ML-based system at least conceptually, developers can share some of the eventual accountability for the outcome of such projects with them.

DP9: Confirmatory Nudging – Use language and representation devices that align well with users´ and managers´ mental models to increase acceptance of the machine learning model

Due to the above-mentioned reliance on domain knowledge in MAN Energy Solutions' aftersales, we decided to integrate our win-propensity prediction system into an existing sales quotation follow-up tool that sales professionals were using daily to perform their tasks. In addition to that, we tried to use a language familiar to the sales professionals whenever we could, for instance, by using speaking terms like "engine type" rather than cryptic terms like "X1, X2" for variables. Here we are using a form of nudging (Thaler and Sunstein 2009) by utilizing a confirmation bias (a tendency to prefer familiar information; Nickerson 1998) to influence the behavior of the sales professionals in a positive way.

DP10: Causality – Choose the machine learning model that aligns best with reality and design it as if it was an explanatory rather than a predictive model to increase model acceptance by users, managers, and developers

Based on the three-gap framework (Kayande et al. 2009), we argue that data scientists should select ML models so that they align as closely as possible to reality. In practice, this means that one should try to optimize the predictive performance of ML models. On the other hand, we

argue that data scientists should design predictive ML models as if they were doing an explanatory study. Explanatory studies focus on causal relationships and are domain knowledge and theory-driven instead of concentrating on data and associations (predictive studies; Shmueli 2010). Moreover, we argue that by designing predictive ML models more like an explanatory study, it is easier for stakeholders to close the gap between their mental models and the ML model, which, based on the three-gap framework, should increase model acceptance.

Furthermore, according to Wheelwright et al. (1998, p. 288), multicollinear variables are not a big issue in purely predictive modeling as they do not affect predictive performance. However, they argue that multicollinearity has to be addressed, 1) if one is interested in the model coefficients, and 2) if one is interested in the marginal effect that an independent variable has on the dependent variable. Informed by this theory, we included a step of hypothesis development in our design method, which represents the development of a causal model that was guiding our sensemaking processes. Moreover, as we are interested in the effects that individual independent variables have on the dependent variable (in the form of Shapley values), we identified multicollinear variables via variance inflation factors analysis and removed or combined them for increased interpretability.

5.4 ADR Study 3: Causal Impact Analysis of Value-based Pricing Strategies

5.4.1 Problem Formulation

Original equipment manufacturers start to servitize business processes and models (Lightfoot et al. 2013). As a part of this transformation towards customer-centricity, they change their pricing strategies from cost-based to value-based approaches (Hinterhuber 2004, 2008; Hinterhuber and Liozu 2014). With value-based approaches, one sets prices in terms of the perceived value that a product brings to the customer instead of using production and logistics costs as the point of departure for price-setting. As a consequence, value-based price-setting strategies require data-driven and micro-level pricing approaches that are tailored to individual material groups. Due to this, pricing analysts have to develop and test many hypotheses that are often only partially backed by experience or sound theories. For this, randomized controlled trials (RCTs; Cochrane

and others 1972) are arguably one of the most rigorous approaches for estimating causal treatment effects. Nevertheless, conducting an RCT for each pricing intervention can be complicated and costly, which makes it infeasible for most companies (Varian 2016). We identified this field problem as a research opportunity to investigate:

How to design causal inference systems that support value-based spare parts pricing decisions?

Our initial solution design was informed by the AI Value Creation Study, which guided our search towards causal inference theory (Hernán et al. 2002; Hernan and Robins 2010; Rubin 1974) in general and quasi-experimental methods (Cook et al. 1979) in particular.

Throughout the cycles of building, intervention, and evaluation, the problem formulation, but especially the latent solution design, was shaped and informed by different theories that helped us explain and/or solve unanticipated problems and solutions. Bartezzaghi and Kalchschmidt (2011) guided and justified, for instance, pre-aggregation related features that we built into the AI system (DP11), Gregor and Benbasat (1999) and Scott and Varian (2013) informed the design of scalability related features (DP12), while Brodersen et al. (2015), Hernán et al. (2002), and (Pearl 1995) informed features of causal modeling (DP13). Moreover, theory on interactive visualizations (Liu et al. 2014) informed our user interface design (DP16), while research on time-series cross-validation (Bergmeir et al. 2018) helped to improve the predictive strength of our AI system.

5.4.2 Building, Intervention, and Evaluation of Solution Artifacts

The concrete design goal was

To construct a system that helps pricing managers and analysts to estimate the causal effects of value-based pricing interventions and by that allow them to test their pricing hypotheses.

After several BIE cycles, we designed and implemented a causal inference system for value-based pricing support. According to Rubin (1974), to estimate causal effects, one should manipulate a treatment variable (e.g., unit price) for a unit of interest, to then compare the actually observed effects of that intervention (e.g., on sales volume), with the potential (counterfactual)

outcome that one would have observed, had the treatment not occurred. Following a quasi-experimental approach (Cook et al. 1979), our system uses a Bayesian structural time-series model (Scott and Varian 2014) to predict the counterfactual outcomes (e.g., in terms of sales volume) of a pricing intervention (unit price change) on a material group (e.g., in terms of sales volume), instead of using an actual control group to simulate counterfactuals. The system then subtracts the counterfactual outcome (Y_0) from the observed outcome (Y_1) to calculate treatment effects. The prediction of the counterfactual outcomes is based on a local-level trend component of the outcomes (e.g., sales volume) of the repriced materials (treated unit), and a number of covariates in the form of, e.g., seasonality terms, but also the outcomes (e.g., sales volume) of other control material groups that had a high pre-intervention correlation with the repriced material group of interest but were not directly affected by the pricing intervention.

Figure 23 provides an overview of the system execution process, Figure 24 illustrates the counterfactual prediction mechanism, and Figure 25 is a snapshot of the user interface.

We continuously evaluated the system throughout the BIE cycles in terms of its functional and structural features but also its usability. Nevertheless, the main evaluation criterion was effectiveness. The key stakeholders received the system very positively. Based on its outputs, the value-based pricing hypotheses of a large repricing initiative of more than 30.000 price changes could be tested. This supported a top-management decision to roll similar value-based pricing strategies out to further regional headquarters and material groups.

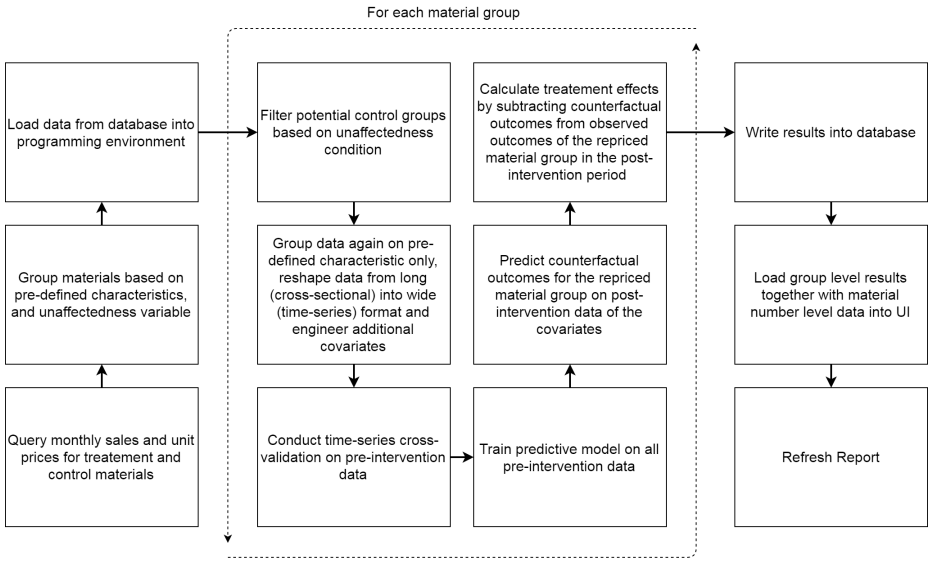

Figure 23. The system execution process

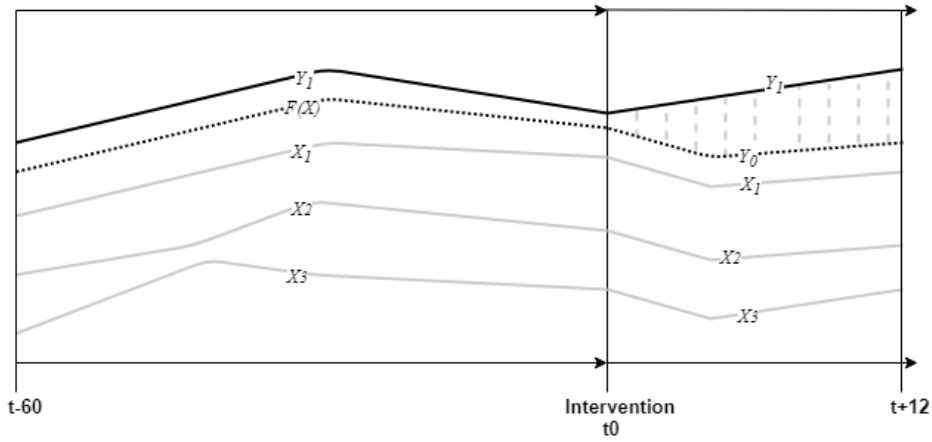

Figure 24. Counterfactual prediction approach (the dashed area represents the treatment effect)

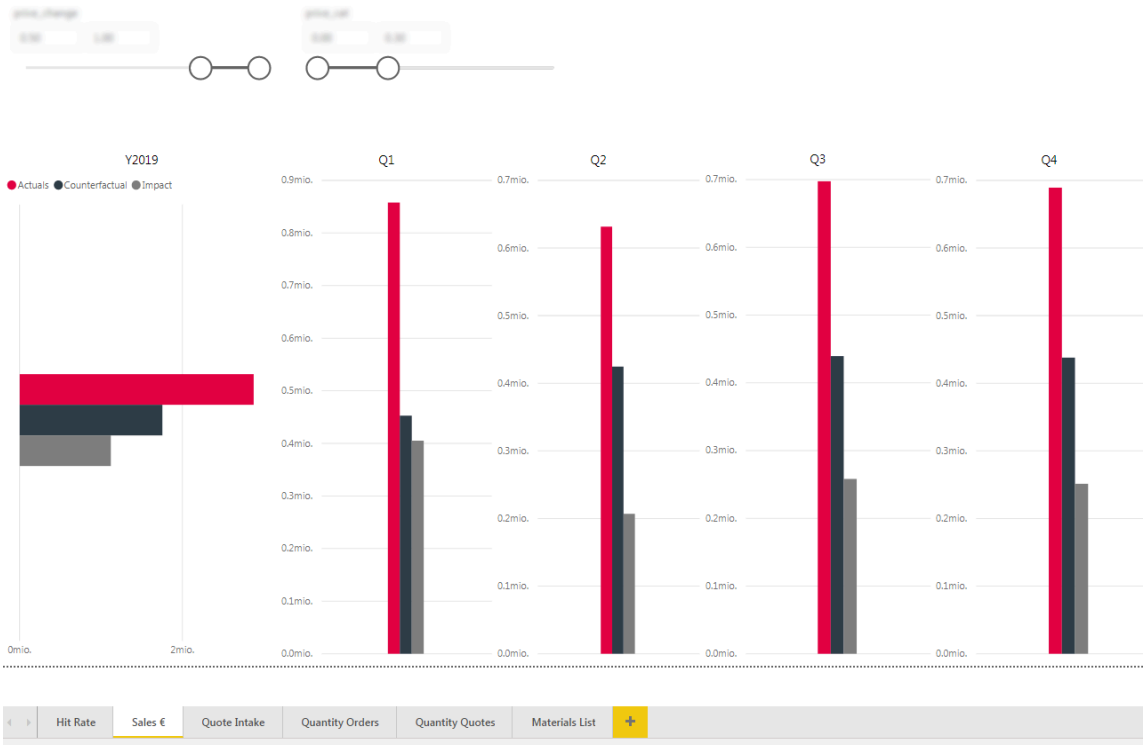

Figure 25. Interactive report with adjustable filters (blurred for confidentiality reasons); red = observed sales revenue, black = counterfactual sales revenue, and grey = the treatment effect

5.4.3 Reflection, Learning, and Formalization of Design Principles

We developed a set of design principles that we abstracted to the class of causal inference systems for value-based pricing support. They may, however, apply to the broader class of AI systems for B2B aftersales support.

DP11: Pre-aggregation – analysts should pre-aggregate lumpy data to improve its predictability

As in other OEM aftersales contexts, we faced the issue of a large and heterogeneous portfolio of spare parts that often showed lumpy demand patterns (Bartezzaghi et al. 1999). In reaction to this, we tried different aggregation approaches that balance predictability and information loss and eventually decided on a monthly unit-price level and price change level based aggregation. At the same time, we consulted the scientific literature (e.g., Bartezzaghi and Kalchschmidt 2011; Zotteri and Kalchschmidt 2007).

DP12: Scalability – analysts should use robust algorithms that rely on few assumptions only and include global explainability features to enable controlled execution at scale

This design principle is addressing the challenge of estimating causal treatment effects for many materials in a time-efficient way. After trying different algorithmic approaches, we eventually decided on using BSTS (Scott and Varian 2014), due to its strengths in time-series modeling and automatic variable selection via spike and slab priors (Scott and Varian 2013). Moreover, during the BIE cycles, we perceived a need for global explainability features to allow us as the developers to keep control over the large-scale execution pipeline. Due to this, we implemented global explainability features such as visual representations of pre-intervention prediction errors, the strongest predictor, and its inclusion probability for a given material group. An approach that we justified by consulting the literature on global explainability (Gregor and Benbasat 1999), and DP8 from ADR Study 2 (Thiess et al. 2020).

DP13: Unaffectedness – analysts should define unaffectedness conditions based on subject-matter knowledge, and causal diagrams and filter model covariates based on them to avoid spillover effects

For causal treatment effect estimation, it is important that the control variables are unaffected by the intervention (Brodersen et al. 2015). At MAN Energy Solutions and in aftersales in general, however, customers buy spare parts in bundles, which means that customers purchase certain parts frequently together. This makes spillover-effects possible. While they could have a high pre-intervention correlation with the treated material group, they could be affected by the intervention. Being inspired by causal graphical modeling (see Hernán et al. 2002 and Pearl 1995), we mapped candidate causes of spillover effects. After having identified a set of potential causes, we considered different approaches to avoid spillover effects. As a result of this, we implemented an unaffectedness condition into the data processing part of the system. This condition assures that only materials that did not have a price change and belong to a different engine type are allowed to be included as a potential control time-series for a given treated material.

DP14: Pre-intervention predictability – analysts should use cross-validation and evaluate treatment effects in light of the pre-intervention predictability to draw more truthful conclusions

At MAN Energy Solutions, we had to model heterogeneous and often lumpy time-series that often showed very different levels of data quality and predictability. This observation, together with research convincingly showing that goodness of fit measures are often insufficient (e.g., Fildes and Makridakis 1995 and Makridakis et al. 1982), inspired us to implement a time-series cross-validation approach (Bergmeir et al. 2018) to calculate pre-intervention mean absolute percentage errors (MAPE). Based on this, we displayed only those treatment effects as significant whose absolute percentage treatment effect was larger than the pre-intervention MAPE. This way, we make sure that treatment effects are not mostly due to differences in predictability.

DP15: Treatment simulation – analysts should add the treatment variable to the model and fix its post-intervention values at its last pre-intervention value to strengthen the counterfactual prediction

At MAN Energy Solutions, we faced, with price changes, a treatment type that usually occurs several times throughout the lifetime of a particular material. Therefore, we had to model the pre-intervention variations that earlier price changes had on the effects of the unit price on aftersales. For estimating the counterfactual outcome time-series, we fixed the unit price to its last pre-intervention value. By doing so, we explicitly utilized the trained model to simulate a situation in which the unit price for the treated material remained unchanged. This further supported the strength of the counterfactual estimation that the overall predictive counterfactual estimation approach (Varian 2016) already possesses by, e.g., incorporating meaningful control time-series (Brodersen et al. 2015).

DP16: Interactive visualization – analysts should create interactive reports instead of static presentations to aid understanding and acceptance

We observed that in the first iteration of the system, users (pricing analysts) were struggling to comprehend some of the inner workings behind our approach. In reaction to this, we designed

an interactive user interface that allows users to explore the results by manipulating different filters, which immediately affects the shape of the displayed diagrams (see figure 25). We justified our design by consulting research into interactive representations of data that suggest that an increase in usability, learning, and understanding increases system acceptance (Liu et al. 2014).

6 Discussion and Conclusion

In the following chapter, I discuss, first, the implications of the IT artifacts (AI systems) in terms of how they enable different AI value creation mechanism, how they support decisions in the B2B aftersales service funnel, and how they work as key enablers for customer centricity and technology-driven service strategies. Moreover, I discuss in what ways they are novel, and therefore, contribute to the knowledge base. Furthermore, I discuss the implications of the developed design principles and theory in terms of how they contribute to AI value creation, and I discuss the development process of design principles with ADR. The chapter closes by discussing some further implications, limitations, and reflections, and overall concluding this PhD project.

6.1 Implications of the IT Artifacts (AI Systems)

6.1.1 Enabling Different AI Value Creation Mechanisms

In the AI Value Creation Study, we identified three AI value creation mechanisms and eight necessary but not sufficient conditions for the different mechanisms. In the following, I discuss how our theory fits with the different AI systems that we built during the three ADR substudies (see Figure 26).

AI System 1 that resulted from *ADR Study 1* is based on our AI system classification, a low-discretion AI task augmentation system, because its value targets were overall improved aftersales service decision making about which customers to contact pro-actively and based on which life-cycle event. While the system's main machine learning outputs are predicted customer lifetime values, the overall end-user facing output was prescriptive. In particular, sales

responsibles got notified in their customer relationship management system that they got a new lead assigned to them. When they opened the lead, they were clearly instructed what to do, e.g., "the ownership of vessel X changed from customer Y to customer Z, please call customer Z for aftersales service opportunities!" The main machine learning type, here, was supervised learning in the form of a hierarchical Bayesian probability model (Platzer and Reutterer 2016) in combination with a gamma-gamma model to calculate the customer lifetime values (Fader and Hardie 2013). Moreover, one can speak of it as a form of optimization. In this case, the decision alternatives were the customers that had an ownership change on a vessel. The customer lifetime values could then be used to score and evaluate the alternatives, based on which, only for customers with high potential customer lifetime values, leads were generated. Moreover, the sales professionals as end-users and decision-makers did only get assigned one "optimal" lead per customer for which they were responsible. While here, human sales professionals were the eventual decision-makers, they had low-discretional free-roam in the decision making process, as the only choice they had left was whether to accept or decline the lead. Also, the decision (call the customer) was implemented by human sales professionals. Here the decision was on a programmable and somewhat structured (whom to call about what?) and operational level.

Also, *AI System 2* that resulted from *ADR Study 2* is, based on our AI system classification, a low-discretion AI task augmentation system. Its value target was better decision making about which aftersales service quotations to follow-up. Also, based on the lightGBM algorithm (Ke et al. 2017), it created predictions following a supervised learning approach. Human sales professionals were making the decision about whom to follow-up based on the win-propensity scores. And the human sales professional also performed the action to call and reach out to a customer to follow up on an open quotation. Also, for this task, the decision problem was structured (programmed), and the decision level was operational.

AI System 3, on the other hand, was a knowledge creation system that follows a data-intensive research approach. Its direct value target is organizational knowing about value-based pricing strategies and their impacts on aftersales. However, indirectly it supports a tactical decision about what particular pricing approaches to use, and in our case, it informed a strategic deci-

sion to roll out a similar approach to another location too. The direct machine learning output of its supervised Bayesian structural time-series model (Scott and Varian 2014) generates predictions of counterfactual outcomes for a repriced material that simulates a situation in which the price change has not occurred. However, the system output is explanatory in the form of treatment effects of price changes on different material groups. Here, the decision-maker in the form of the pricing manager was human, and also the action of price setting was still a human responsibility. In contrast to *AI systems 1* and *2*, however, the decision making task is rather non-programmable and unstructured.

Overall, we could successfully implement the systems because all the necessary conditions were met. The practitioners whom we collaborated with had *deep domain knowledge*, and we could bring in *strong data science capabilities*. Also, MAN Energy Solutions has a *mature data infrastructure* around a well-developed and maintained aftersales services data warehouse. Further, the practitioners supported us in integrating our systems into the IT infrastructure (*DevOps*). When it comes to *UX capabilities,* the practitioners helped us in technically implementing our often user acceptance and interpretability related design features.

Both *AI System 1* and *AI System 2* could shift towards a process automation mechanism; here, both systems would be candidates for process exploitation systems. For *AI System 1,* one could instead of calling the customer in person, send an email automatically: However, this would arguably constitute a lower service level. Following (Huang and Rust 2017), one could approach customers with a low potential customer lifetime value with a process automation system while approaching customers with high potential customer lifetime value with the current task augmentation approach. Also, for *AI System 1,* the environment around the data-driven lead generation is stable, legal, and ethical constraints are not obvious. For *AI System 2*, one could follow-up on open quotations with low propensity scores, e.g., via automatically sending emails or default promotions, while still following up on the quotations with high- and middle-propensity scores quotations in person. Also, here, stable environments and legal and ethical constraints are no obvious issues. For *AI System 3,* shifting towards task augmentation and process automation is not easily possible. One could, however, based on the tested value-based

pricing hypotheses, design an AI system that, based on learnings, automatically prices spare parts. However, during all ADR sub-studies, we could not integrate our systems with the transactional systems and processes (ERP systems), which means that process automation is generally not possible until the IT department provides flexible interfaces.

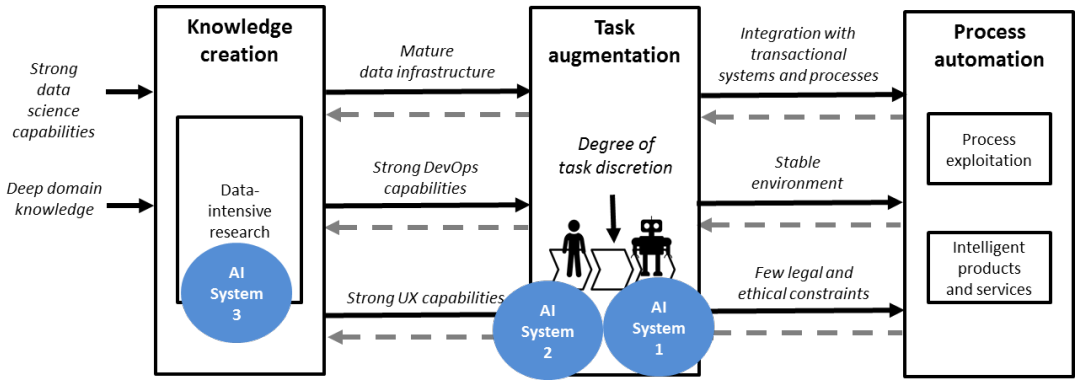

Figure 26. AI Systems 1, 2, and 3 related to the AI value creation mechanisms and systems

6.1.2 Decision Support in the B2B Aftersales Service Funnel

All of the three AI systems, which we developed during the ADR program, address different phases in the B2B aftersales service funnel (see figure 27).

AI System 1, the *Data-driven Lead Generation*, addresses the pro-active creation of aftersales leads. Moreover, *AI System 1* supports the conversion from leads to opportunities (quantifiable business options). This support is enabled via the high quality of lead events (e.g., shipowner changes), the selection of leads for high-potential customers only, and the prescriptive formulation of how to take action on such leads.

AI System 2, on the other hand, directly addresses the conversion of aftersales quotations into aftersales orders. In particular, it supports sales professionals in the decision about which open quotations they should follow up. This makes resource allocation processes much more efficient, and can also lead to increased conversion rates because, following a customer-centric approach (Fader 2012), sales professionals can now focus on the high-potential quotations and customers, instead of spending most of their time on quotations that have a very low win pro-

pensity, e.g., quotations of customers who only want to get an overview of prices, but are not actually interested in a purchase.

AI System 3, again, addresses the conversion of aftersales quotations into aftersales orders as well as it supports pricing managers' validation of their value-based pricing strategies, which enables better and more personalized prices, e.g., higher prices for parts with high perceived customer value, but lower prices for many of the parts with low perceived customer value (Hinterhuber 2004, 2008; Hinterhuber and Liozu 2014). This, in turn, can increase the conversion rate. Moreover, more attractive prices can create pull-mechanisms that, in turn, can lead to requests for aftersales quotations.

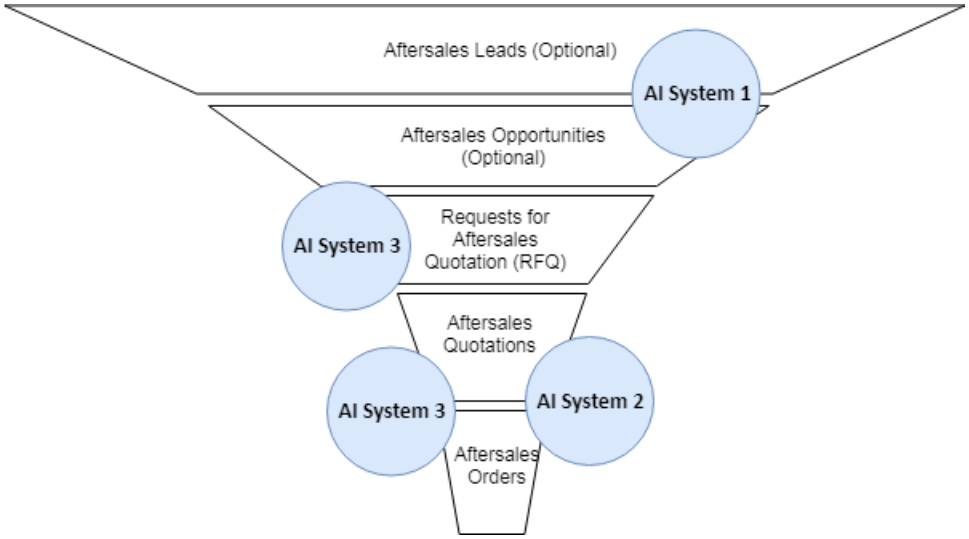

Figure 27. AI systems in the B2B aftersales service funnel

6.1.3 Key Enablers for Customer Centricity and Technology-Driven Service Strategies

Huang and Rust (2017) propose a typology of technology-driven service strategies (see Figure 28). The horizontal axis in their typology represents a range of service attributes from standardized (left) to personalized (right). The vertical axis represents a range of customer attributes from transactional (at the bottom) to relational (on top). Transactional customers are almost exclusively interacting with the company via e-commerce interfaces. Relational customers, on the other hand, interact with frontline professionals too. Moreover, standardized

services are or should be offered for customers with homogenous demand patterns, while customers with heterogeneous demand should be approached personalized. Also, they argue, that transactional strategies are better when the potential customer lifetime value of a customer is low, while relational strategies are better when the potential customer lifetime value of a customer is high.

Based on this, Huang and Rust (2017) argue that relational and standardized strategies can be built around databases and customer relationship management systems. Transactional and standardized strategies, on the other hand, they argue, can be based on automation technology, e.g., robotic process automation. Moreover, personalized and transactional strategies can be based on big data analytics or what we call knowledge creation or data-intensive research systems that, for instance, use unsupervised clustering machine learning algorithms. Finally, according to them, AI systems are the key enabler for personalized and relational service strategies.

The typology supports our findings and helps to embed our contributions in a larger strategic context. At the same time, our contributions validate the typology. We built all three AI systems based on aftersales transactional data that, in general, is highly heterogenous and "lumpy" (Bartezzaghi et al. 1999; Bartezzaghi and Kalchschmidt 2011). Following this, the typology would suggest following a personalized strategy.

AI System 1, the *Data-driven Lead Generation*, for instance, explicitly involves a step to calculate customer lifetime values accurately with an advanced Bayesian probabilistic machine learning model (see Fader and Hardie 2013; Platzer and Reutterer 2016). Also, with shipowner changes, we generate leads based on events that are highly customer-lifecycle dependent and, therefore, personalized. Moreover, based on the customer lifetime values, the system can filter the leads so that only leads of customers with high potential customer lifetime values were selected. Such an approach is also fully aligned with the concept of customer-centricity (Ascarza et al. 2017; Fader 2012), in which one first identifies customers with high-potential customer lifetime values, and

then to focusses on serving them on a high level of quality, instead of serving every customer at a mediocre level.

Huang and Rust (2017) suggest that IT and big data analytics allow approaching the high-potential customers on a relational level with frontline professionals, while also serving the low-potential customers, e.g., via e-commerce interfaces. Following this suggestion, one could develop *AI System 1* further by, e.g., still sending the leads for high-potential customers to the frontline sales professionals, but using the leads for low-potential customers, for instance, as the basis for targeted advertising by automatically sending emails to them.

A similar approach would be possible for *AI System 2*. While it does not explicitly predict customer lifetime values, its win-propensity scores are based on transactional customer data too. Due to this, an open quotation from a customer that would have a high customer lifetime value, would, in general, also have a rather high propensity score. And quotations from a low-customer lifetime value customer would generally also be low. For *AI System 2*, the differentiated approach would be easier, since, at the point of prediction, quotations are already in the system, and one could send automatic follow-up emails for low-win-propensity quotations and contact the customer in person for high win-propensity quotations.

AI System 3, in turn, helps to estimate the effects that value-based pricing strategies have on sales. In value-based pricing, one of the key assumptions is to price materials differentiated based on, for instance, customer characteristics and their perception of how valuable a material is to them (Hinterhuber 2004, 2008; Hinterhuber and Liozu 2014). In that sense, the system supports more personalized service strategies, and one could potentially use it to evaluate the effects of, for instance, setting up long-term price agreements with high potential customers. For low potential customers, the system could estimate the effects of assigning prices to them without the possibility of such special agreements.

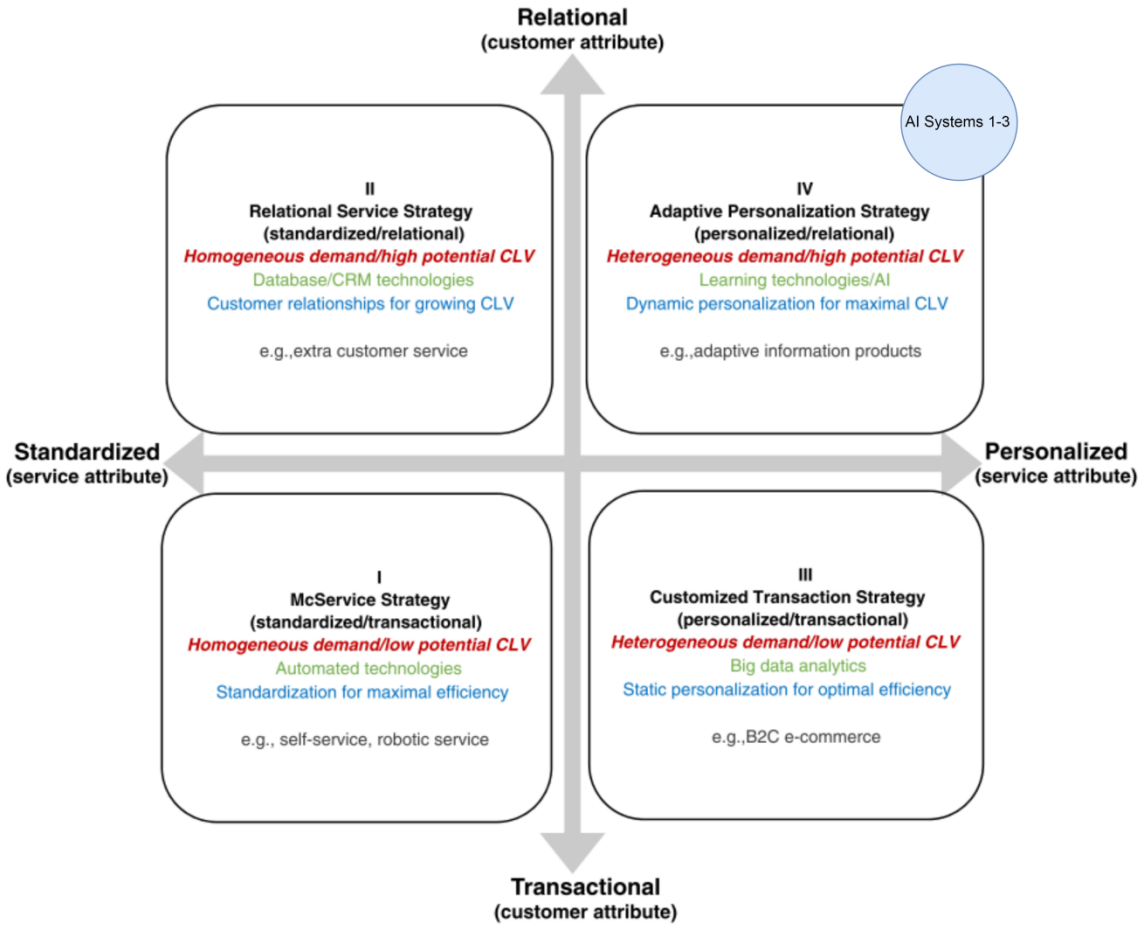

Figure 28. AI systems in the Technology-driven Service Strategy Positioning Map (Huang and Rust 2017)

6.1.4 Artifact Novelty

According to Davis (2005), a design science based PhD thesis can contribute to knowledge when it "develops and demonstrates new or improved designs of a conceptual or physical artifact. […] The contribution may be demonstrated by reasoning, proof of concept, proof of value-added, or proof of acceptance and use" (p. 18).

Gregor and Hevner (2013) developed a design science knowledge contribution framework (Figure 29) that differentiates four different quadrants. Inventions are artifacts that represent the first solutions for new and unexplored problem contexts. Artifacts here are mostly on the first level of theoretical contributions (situated implementations of artifacts; Gregor and Hevner 2013). The improvement quadrant refers to the development of novel solution artifacts to well-known problem contexts. With this contribution type, researchers need to convincingly argue in

what way their artifact is better than the existing solution artifacts. Exaptation of existing solutions artifacts to unexplored problem contexts is a clear knowledge contribution. However, under the constraint that, in the unexplored problem contexts, one needs to deal with challenges that have not been present in the well-known problem context for which the solution artifact was originally designed. Routine designs, on the other hand, do not constitute a great research opportunity as they concern the application of known solution artifacts to known problem contexts (Gregor and Hevner 2013).

AI System 1 that resulted from *ADR Study 1* was quite clearly an *improvement* as we developed with the data-driven lead generation a new solution for the known problem of lead generation, even though, the lead generation problem was new for MAN Energy Solutions who were using more reactive aftersales service approaches before.

AI System 2 that resulted from *ADR Study 2* was in-between the continuum of *improvement* and *exaptation*. With our sales win-propensity prediction system, to some degree, we extended the known solutions (e.g., Bohanec, Kljajić Borštnar, et al. 2017; Yan, Gong, et al. 2015) in win-propensity scoring for leads and opportunities to win-propensity scoring for quotations. However, our system was also an *improvement* as none of the other documented approaches used an algorithm that was theoretically as efficient as the lightGBM algorithm that we used (Ke et al. 2017) or which had model explanations of the same theoretical quality as our Shapley value-based approach had (Scott M. Lundberg et al. 2018; Lundberg and S.-I. Lee 2017; Lundberg and S. I. Lee 2017), nor did they model quotation age (or opportunity age) explicitly with a second-level conditional probability model.

AI System 3 that resulted from *ADR Study 3* Was an *exaptation* of the prediction based counterfactual estimation approach that Varian (Varian 2016) proposed, and Brodersen et al. (2015) developed. In particular, we extended the solution known from digital B2C marketing to value-based pricing support to a B2B aftersales context (see DP11 and DP12). But we also improved the solution in general by incorporating an unaffectedness condition (DP13), a measure of pre-intervention predictability (DP14), and features of treatment visualization (DP15).

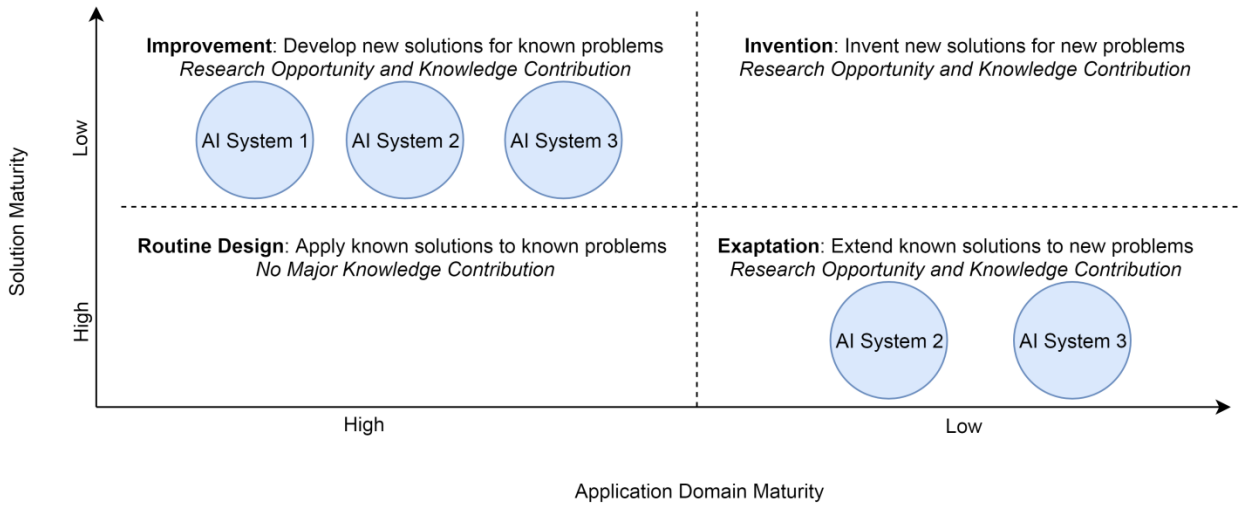

Figure 29. AI Systems 1-3 in the DSR Knowledge Contribution Framework (Gregor and Hevner 2013)

6.2 Implications of the Developed Design Principles and Theory

In the following section, I discuss how the different design principles relate to each other and what kind of theoretical contribution they are.

Here, I want to note that, in each of the ADR studies, they were addressing slightly different solution classes. We formulated design principles 1-4 from *ADR Study 1*, for the solution class of data-driven decision making. Design principles 5-10, in turn, we formulated for the solution class of explainable sales win-propensity prediction systems and machine learning systems in general, and design principles 11-16, we formulated for the class of causal inference systems for value-based spare parts pricing. Nevertheless, they all belong, too, to the common class of AI systems for B2B aftersales decision support.

6.2.1 Design Principles in the AI Value Creation Process

In the following, I discuss which stages of our *AI value creation process* the different design principles address (see Figure 30 and Table 2).

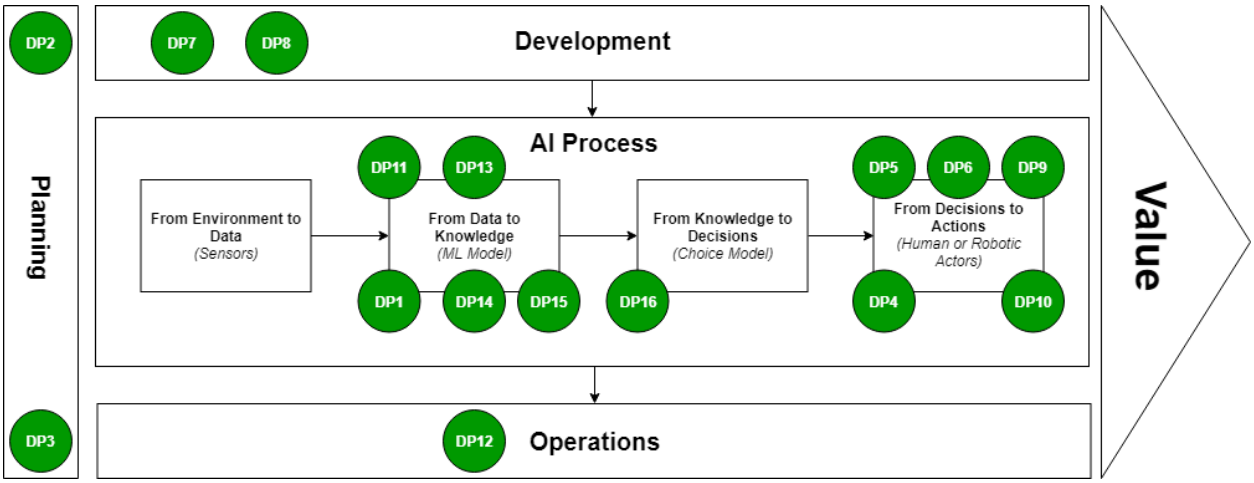

Figure 30. Design Principles in the AI Value Creation Process

In the planning phase, it is crucial to get management sign-off for AI system implementation projects. DP2, *Comprehensibility*, and DP3, *domain knowledge*, address this issue of management approval by helping to increase managers' acceptance of a system.

Also, in the development stage, managers must understand how AI systems work to avoid unaccountable AI or AI systems for which only the developers are outcome accountable, as otherwise, increased stress levels for data scientists, opportunism, and escalation of commitment to failing courses of actions can follow (see Section 2.3.2). DP7, *Accountability*, and DP8, *Global Explainability* address this issue.

Turning data into knowledge is the core of machine learning and, thus, AI systems. In this substage of an AI process, the accuracy of AI predictions has a significant impact on how usable its generated knowledge in decision making for the end-users of an AI system is. The design principles DP1, *Theory-driven Modeling*, DP11, *Pre-aggregation*, DP13, *Unaffectedness*, DP14, *Pre-intervention Predictability*, and DP15, *Treatment Simulation* address this.

The *Knowledge to Decisions* stage is all about using knowledge effectively to make high-quality decisions. Here, DP16, *Interactive Visualization*, addresses the issue.

The *Decision to the Action* stage is concerned with implementing decisions as effective actions. It is here where AI acceptance related problems such as algorithm aversion surface. DP4, *Actiona-*

bility, DP5, *Local-contrastive Explainability*, DP6, *Selective Visualization*, DP9, *Confirmatory Nudging*, and DP10, *Causality,* address these issues.

The operations stage is all about ensuring a high-performing and continuous execution of the AI system. DP12, *Scalability*, addresses this.

Table 2. Design Principles in the AI Value Creation Process and the kinds of Actors that they directly Affect

AI Value Creation Process	**Design Principles**	**Directly affected actor**
Planning	**DP2:** Comprehensibility **DP3:** Domain Knowledge	Manager
Development	**DP7:** Accountability **DP8:** Global Explainability	Manager
From Data to Knowledge	**DP1:** Theory-driven modeling **DP11:** Pre-aggregation **DP13:** Unaffectedness **DP14:** Pre-intervention predictability **DP15:** Treatment simulation	End-user (decision maker)
From Knowledge to Decisions	**DP16:** Interactive visualization	End-user (decision maker)
From Decision to Actions	**DP4:** Actionability **DP5:** Local-contrastive explainability **DP6:** Selective visualization **DP9:** Confirmatory nudging **DP10:** Causality	End-user (decision maker)
Operations	**DP12:** Scalability	Developer (data scientist)

6.2.2 Development of Design Principles with ADR

During the three ADR studies, we followed the ADR method closely. However, as a method in use, people interpret it and apply it differently (Haj-Bolouri et al. 2017). In the following, I discuss how we developed design theory with ADR during this PhD project. To support this discussion visually, I illustrate it in Figure 31. In the figure, I distinguish three levels of abstraction, first, the Observable Practice Level, second the Latent Design Theory Level, and, third, the Kernel and Justificatory Theory Level.

In our case, relevant field problems at MAN Energy Solutions inspired and informed our search for literature that helped to structure and formulate the problem context of B2B aftersales decision support. This problem formulation, again, informed the search for theory that could help to solve the problem, in our case, the body of knowledge on AI and customer-centric and AI-driven services. This initial theory ingrained solution design, informed the building, intervention, and (formative) evaluation of artifacts (AI systems) at MAN Energy Solutions, which, again, shaped the eventual operational IT ensemble artifact, in our case the implemented AI systems. From here, one could conduct a behavioral summative analysis of the ensemble artifact to extend or validate the theory that was initially inscribed into the artifact before it was shaped by its organizational context. We, however, focused on abstracting the situated knowledge of the ensemble IT artifact to a larger class of solutions (AI systems) by formalizing the design principles and methods that it was based on. At the same time, the reflection upon the eventual shape of the artifact informs the search for theory that can explain and justify why its features, principles, and methods were effective in helping to solve the problem.

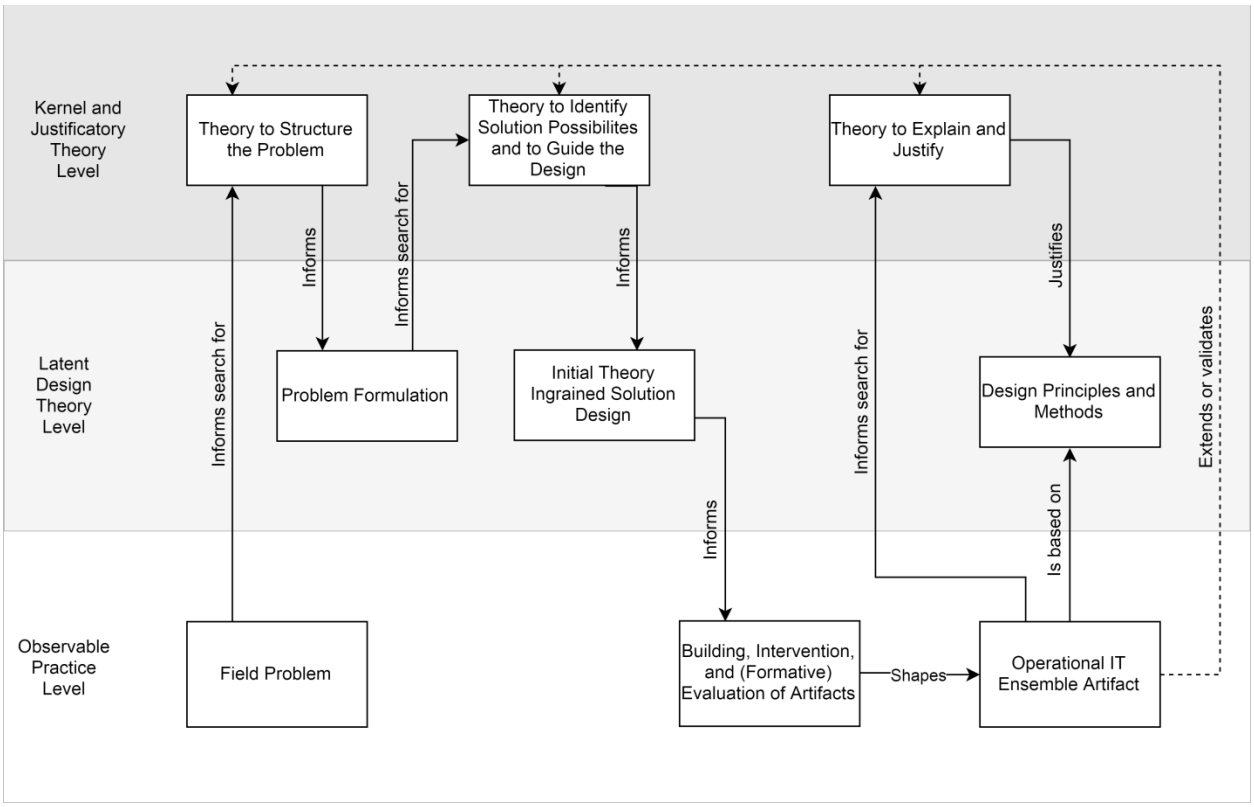

Figure 31. Developing Design Principles and Methods with ADR

6.3 Further Implications

This PhD project is one of the first that used ADR in a larger ADR program consisting of different ADR sub-studies conducted in the same organizational context. It shows how to use such different ADR studies to create design theory about an overall class of systems even though the sub-classes may be different.

Moreover, the PhD project shows how one can use design science research methods to make contributions to decision support systems more rigorous. According to Arnott and Pervan (2014), 88% of all design science papers about decision support systems in their sample did not explicitly describe a research method. Moreover, they found that only 14% of the designed artifacts were used in field contexts.

Based on Gregor and Jones (2007) and Walls et al. (1992), I argued in Section 4.3 that a full-blown design theory should consist of at least:

(1) A description of the goals and the class of systems that the design theory aims at

(2) Design Principles in some form

(3) Falsifiable propositional statements

(4) Justificatory knowledge or kernel theory

(5) The description of an implementation method

First, this PhD thesis, for all three AI systems, describes the design goals, and the overall class of systems that the design theory aims at. Second, for all three AI systems, it formalizes design principles, third, in a falsifiable way. Fourth, for all three AI systems, this PhD thesis discusses kernel and justificatory theory components. Fifth, for AI System 1 and 2, it presents an implementation method. Following this, I argue that both ADR studies 1 and 2 actually are a more developed form of design theory that goes beyond situated implementations of IT artifacts and nascent design theory (see Gregor and Hevner 2013; Gregor and Jones 2007).

6.4 Limitations and Reflections

Despite the discussed approaches to generalize design knowledge, it is arguably still more difficult to generalize it than it is to generalize knowledge from large quantitative studies of many subjects. Due to this, the findings and design principles that are some of the main contributions of this PhD project may, in some cases, not necessarily be generalizable but at least transferable to comparable contexts.

Moreover, ADR is still an evolving method. Because of this, it can be difficult to assess such studies as a reviewer but also conduct and publish them as a researcher. I experienced that papers got rejected due to the sometimes confusing structure that can occur if one has to introduce many concepts such as the ADR method, the scientific motivation, the practical motivation, a general theoretical background, justificatory knowledge, artifact descriptions, discussion of design principles, etc. Moreover, some people seem to be confused about the role of evaluation in ADR. While I had enough space to discuss this aspect in detail in this PhD thesis, I sometimes struggled to do so in the different papers. Especially in my setup in which I was the main re-

searcher and lead developer at the same time, I would have sometimes liked to focus more on either one of the two (being a researcher or being a developer).

Moreover, I experienced high outcome uncertainty due to the application of advanced AI technology but also with regard to organizational commitment. I invested, for instance, a lot of time in 2 projects that eventually were discontinued for several reasons, amongst them unclear responsibilities as normally the lead developers are also responsible for the system operation. Another issue was that sometimes the organizational change, and especially the momentum just changed. Moreover, I was involved in one project about customer segmentation that, simply, did not turn out to be theoretically interesting enough to publish, even though I invested a lot of time into it.

6.5 Conclusion

This PhD project contributes to information systems, AI value creation, and industrial marketing research by investigating the question *How to design AI systems that support customer-centric aftersales processes and strategies?*

In particular, this PhD project contributes with design principles and theory developed from the situated implementation of three novel AI systems that are informed by state-of-the-art knowledge about AI value creation and address the particular challenges of the B2B aftersales context. Moreover, the design principles guide designers in constructing artifacts of the class of AI systems for B2B aftersales decision support, and where applicable, the broader class of AI decision support systems.

With these contributions, I have achieved the research goals, and, therefore, answered the main research question of this PhD project. Moreover, this PhD project answers calls for more scientific contributions on applications of AI in industrial marketing contexts (Martínez-López and Casillas 2013). And by proposing how to "[establish] procedures for data […] analysis to improve decision-making" (Mora Cortez and Johnston 2017, p. 9) in industrial marketing contexts, it shows how to "resolve real problems that B2B marketers will face during the next three to five years" (p. 6).

Furthermore, this PhD project fills a gap in the literature on AI value creation by investigating *How can organizations create and sustain value through applications of AI?*

In particular, this PhD project conducted a qualitative study that investigated how real-world designers approach the holistic data-to-insights-to-decision-to-actions-to-value path with organizational applications of AI. The study focuses on the role that organizational structures and human decision making processes play in the design and implementation of AI and analytics systems and how such systems shape their organizational environment. Based on this, it suggests how to circumvent challenges not only in converting data to insights, but in converting such insights into decisions, and such decisions into actions and value (see Sharma et al. 2014).

Part B

1 Paper I

Challenges and Enablers along the AI Value Creation Process

By Oliver Müller,

Tiemo Thiess,

Konstantin Hopf,

and

Arisa Shollo

Abstract

Recent years have seen major technological breakthroughs in the field of machine learning (ML), and corporate investments in artificial intelligence (AI) are expected to reach nearly $98 B in 2023. However, AI's impact on the productivity of organizations and economies has been modest so far. Many ML projects never leave the pilot phase, and companies have difficulties extracting measurable value from their AI initiatives. To shed light on this paradox, we studied more than 50 projects implementing machine learning in organizations from different industries and of varying sizes. Based on our research, we conceptualize the process of creating value from AI, identify major challenges along the way, and propose enablers for overcoming these challenges. Our findings can inform CIOs by giving guidance for planning, running, and profiting from AI projects.

Keywords: Artificial Intelligence, Organizations, Implementation, Challenges, Enablers

1.1 The History of AI Automation

Since the 1950s, more and more work processes have been automated by information technology (IT), with machine learning (ML) currently being the most advanced automation technology. The push towards automation started with rule-based systems that consisted of basic hard-coded what-if conditions. These so-called expert systems started the first wave of AI. Due to their rule-based nature, these systems were especially suitable for automating clearly structured routine work processes that leave workers low judgemental discretion. Since the early 2000s, ML technology has changed the AI landscape and initiated the second wave of AI automation. In contrast to rule-based AI systems, ML-based AI systems do not rely on human crafted rules to perform a task, they inductively discover associative relationships among variables in large sets of data, and thereby learn the rules themselves. For example, while the ELIZA chatbot (Weizenbaum 1983), a famous rule-based AI system developed in the 1960s by the MIT Artificial Intelligence Laboratory, used a hand-crafted rule and knowledge base, modern conversational agents like Amazon Alexa are largely based on machine learning techniques and, therefore, able to learn from experience and continuously improve over time. The differences between rule-based and ML-based AI systems are visually summarized in Figure 1.

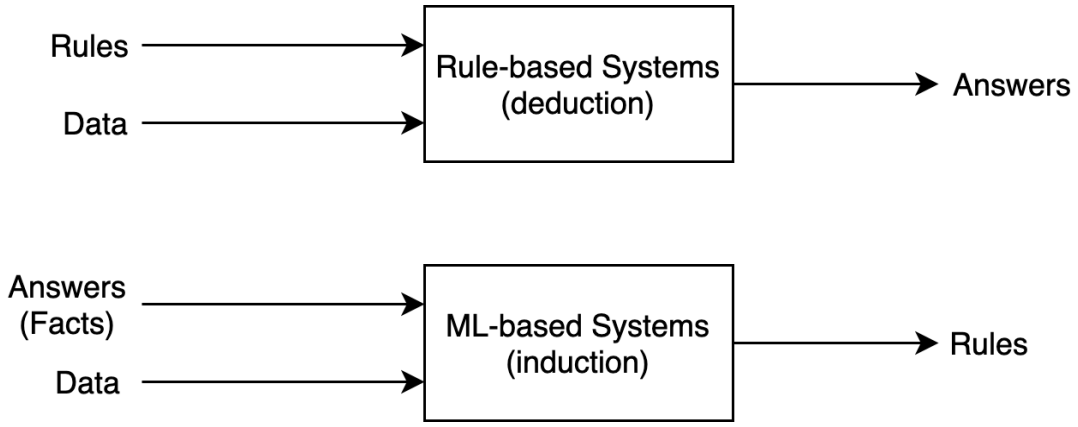

Figure 1: Difference between rule-based and ML-based AI systems (François Chollet 2017)

The first wave of AI was not aimed at replacing human knowledge workers, but rather at augmenting knowledge work by improving humans' information processing capabilities (e.g., attention span, processing speed, data storage, and retrieval). As second-wave AI systems are able to learn the rules themselves, instead of being programmed by humans, they allow for the

automation of high discretional knowledge work, such as the creation and application of knowledge in decision making processes. For example, ML-based AI systems have outperformed human experts in tasks ranging from game playing (Perrault et al. 2019) (e.g., Jeopardy!, Go, Starcraft) to trading stocks (Patel et al. 2015) or diagnosing diseases (McKinney et al. 2020). Driven by these success stories, spendings on AI systems are skyrocketing and expected to reach nearly $98 B in 2023 (International Data Corporation 2019). However AI's impact on the productivity of organizations and economies has been modest so far (Brynjolfsson et al. 2017). Recent studies (Fountaine et al. 2019) report that organizations, besides a handful of unicorns, have difficulties extracting value from their AI initiatives; it seems that most AI projects never leave the pilot phase.

With the goal to pinpoint the challenges companies are experiencing, and to identify enablers for overcoming these challenges, we have studied more than 50 projects implementing machine learning in organizations from different industries and of varying size. While we selected various AI professionals for our expert interviews (e.g. data scientists, data engineer, data science department manager, CEO of AI startup, digital innovation manager), we primarily focused on data scientists as interview partners, given that they are responsible to set up AI and are involved in the whole project lifecycle. Our data collection was separated in two rounds. In the first round (between October 2018 and January 2020), we conducted 40 semi-structured interviews with data scientists (some interviews covered more than one project). Interviews followed an interview guide that focused on concrete AI projects in which the interview partner had been involved in the past, or in which they were currently involved. All interviews were recorded and transcribed; the interviews had a mean duration of 50.25 minutes (10 minutes standard deviation), resulting in 620 transcribed pages. The data from this first phase formed the primary foundation for our data analysis and development of findings. A second round of data collection (in January and February 2020) focused on the further validation of our findings. For this, we organized a workshop with data scientists on January 16 with 13 participants (seven of them were interview partners, while six were new participants). In this focus group, we presented our findings and asked for feedback.

1.2 The AI Value Creation Process

To guide our data collection and analysis and systematically summarise our findings, we structured our study along the AI value creation process. This process model—depicted in Figure 2—suggests that for AI value creation, four main elements are essential: the core AI process as well as the planning, development, and operations of the core AI process.

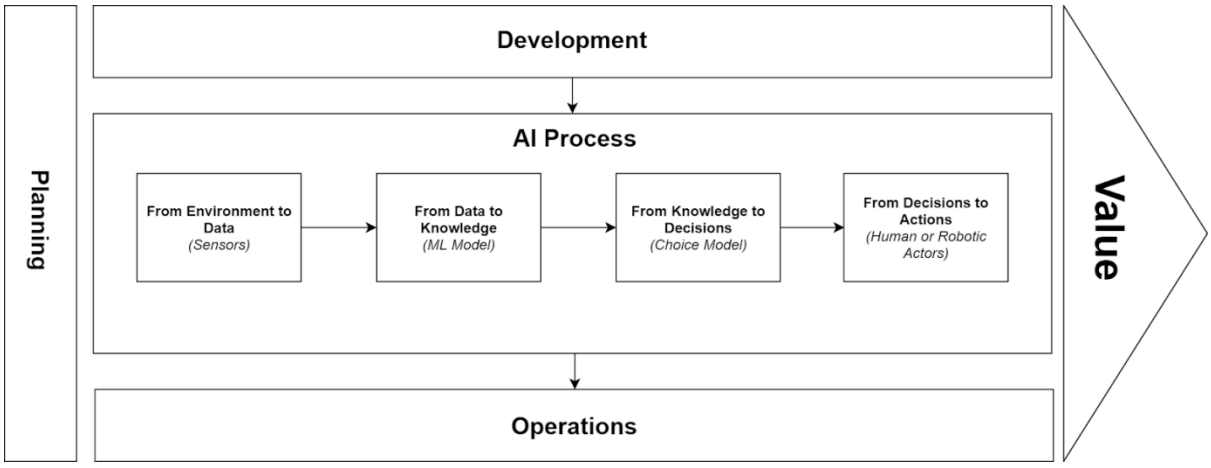

Figure 2: The Artificial Intelligence Value Creation Process

The **AI process** is based on the augmented or automated execution of work activities through ML-based systems. It consists of four main steps: from environment to data, from data to knowledge, from knowledge to decisions, and from decisions to actions.

To transmit physical observations **from environment to data,** one can of course utilize sensors, including cameras and microphones, that can digitize signals like light, noise, or temperature. However, in brick-and-mortar businesses it is more common that front-line workers or customers themselves enter information manually into forms, files, or systems. In digitally born companies, on the other hand, many business interactions are mediated through IT, which means that sensing or observation activities happen as a by-product of daily business. In all cases, the digitization process usually generates noisy data and humans need to curate, clean, and transform data before it can be digested by machine learning algorithms.

To transform **data to knowledge**, AI processes utilize machine learning algorithms that inductively identify trends and patterns in the data and learn rules on how to predict outcomes based on a given set of input variables. A classical example is the prediction of customer churn. A supervised ML model could, for instance, identify the pattern that for each month a customer is not using a service he or she has subscribed to, the likelihood that the customer will cancel the subscription within the next six months rises by five percent. While such models clearly enhance humans' information processing capabilities, they require intensive human involvement and supervision to assure the quality and correctness of the knowledge created.

The very core mechanism of AI value creation is decision making. To transform **knowledge to decisions**, AI processes can either automate decision making or augment decision making. In the automation case, decisions are made by translating the scores or probabilities computed by a prediction model into decisions. In the simplest case, this comes down to setting a threshold for rounding scores to discrete decisions. For example, if the calculated churn score is above 50%, the customer is considered to be inactive. When the decision task allows several alternative courses of action to be taken, utility functions can be applied that find the most optimal alternative in terms of expected utility. For instance, a sales representative with limited time available might be faced with the problem of choosing which customer with a high churn probability to contact first. Here, a utility function can help to choose the most optimal alternative, for example, by combining the computed probability of churning with an estimated customer lifetime value to calculate the future expected value of the customer. In contrast to the decision automation case, in the decision augmentation case a human agent is making the final decision about what course of action to take and uses the knowledge captured in the ML model only as one input factor of many. An example would be a radiologist who uses an AI system that visually highlights certain regions on an MRI scan which might contain benign cells. It is unlikely that the radiologist bases her decision solely on the outputs of the AI system; instead she will probably consult other information sources to form and test different hypotheses about a potential diagnosis.

To transform **decisions to actions**, AI processes rely on human or artificial agents. An example of human agency would be a scenario in which a sales representative, triggered by a decision made or augmented by an AI system, would call a customer who has been identified as inactive. If, in contrast, the AI system would autonomously act by sending out a promotional email to the customer with the goal of reactivating him, we would speak of artificial agency. Such an advanced form of automation, ranging all the way from data capture to action taking, requires a complete digitalization of all interactions and transactions between the system and its environment. Such completely digital feedback loops are possible in some technical systems (e.g., self-driving cars, autonomous robots) and online or mobile business models (e.g., online social networks, e-commerce websites). In the radiology case, and many other knowledge work settings, this level of digitalization has not yet been realized and, therefore, a human in the loop is still required to carry out the final action.

Planning describes all activities related to the problem definition (e.g., by diagnosing a deviation from a plan or by identifying new opportunities) and the design of AI solutions (e.g., assessing whether the problem can be solved via AI and which design is most suitable for a given context). As indicated, the planning phase is continuous, emphasizing the need for adaptable approaches when managing AI projects.

Development has close feedback loops to the planning element and refers to those activities related to actually building IT artifacts in the form of, for example, databases, ML models, choice models, user interfaces, and actuators. In the initial stages of AI projects, ML models are often built as prototypes that are necessary to get management approval and secure resources for conducting the actual project.

Operations emphasizes the fact that AI processes are never "finished" in the way tangible products are. They need to be continuously operated, improved, and maintained by AI professionals, such as, data scientists and data engineers. AI operations also include the execution of support processes for monitoring predictive performance and the real-world impact of the main AI process.

1.3 Challenges and Enablers in the AI Value Creation Process

Of course, merely following the above outlined process does not guarantee value creation. In fact, our in-depth interview study, which involved investigating more than 50 independent AI projects (see Appendix for an overview of projects), shows that firms are experiencing numerous barriers along the process. In the following, we present seven core challenges that organizations and AI professionals face in their pursuit of creating value with AI. Besides describing the challenges, we also describe mitigation strategies (enablers) to overcome them. Figure 3 and Table 2 summarize the challenges and enablers and map them to the phases of the AI value creation process.

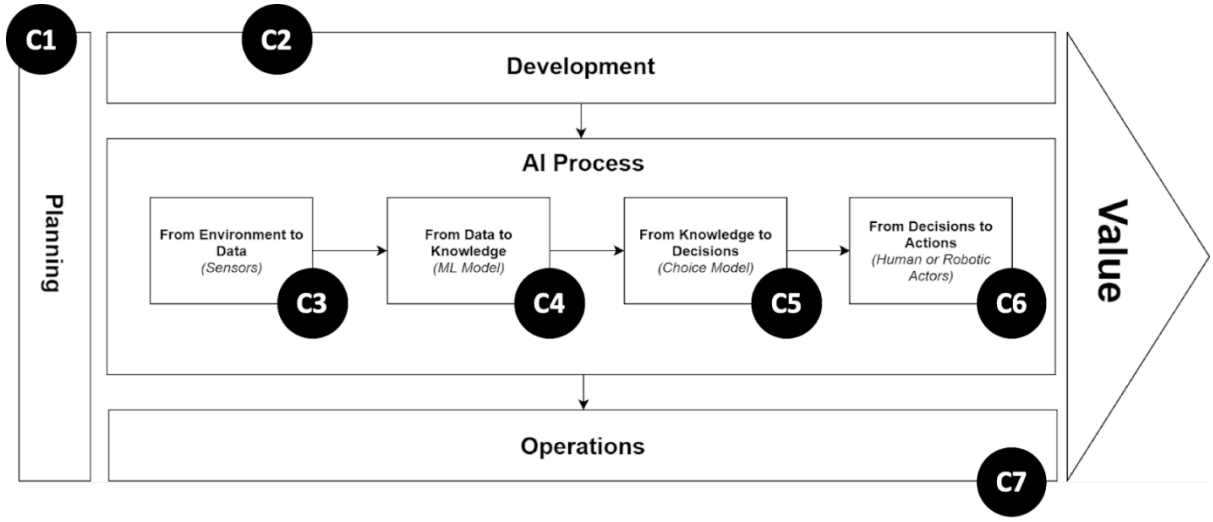

Figure 3: Mapping of Challenges to the AI Value Creation Process

1.3.1 Challenge 1: Inflated Management Expectations

Recent years have seen breakthroughs in the field of AI, both in artificial laboratory settings and real-life applications (Perrault et al. 2019). These success stories are being prominently covered in the news and drive investments in enterprise AI systems. As the number of news stories and amount of dollars spent are rising, managers' expectations of the capabilities of AI are rising, too. Many of our interviewees reported that their departments are overrun by requests for projects. And many of these requests contain unrealistic expectations, which we identified as a major challenge in the planning phase of an AI project. It just seems that nowadays, *"a lot of inves-*

tors and board members expect to have AI in the process", as one of the interviewed data scientists from a media company stated.

However, we are far from artificial general intelligence, machines cannot do the full range of tasks that humans can do. One enabler to overcome the challenge of inflated management expectations is to conduct a reality check on what machine learning can and cannot do. Researchers from MIT and Carnegie Mellon University developed a simple questionnaire for determining a task's suitability for machine learning (see Figure 4). Tasks that are especially suited for automation via machine learning are characterized, among others, by the existence of well-defined inputs and outputs; large digital datasets describing the inputs and outputs; and clear feedback, goals, and metrics. Tasks like image-based medical diagnosis or stock trading fit well to this profile; however, there are also plenty of knowledge work tasks that are not suitable for automation through today's AI systems, such as, creative design, communicating and interaction with customers, or caring for patients.

What makes a task suitable for machine learning?

1. The task involves transforming well-defined inputs to well-defined outputs.
2. Large datasets containing input-output pairs exist.
3. The task has clear goals and metrics and provides clear feedback.
4. No long chains of reasoning that require background knowledge of common sense.
5. No need for explanations of how a prediction was made.
6. Tolerance for error and no need for provably correct solutions.
7. The function being learned should not change rapidly over time.
8. No specialized dexterity, physical skills, or mobility required.

Figure 4: Checklist for assessing a task's suitability for machine learning (SML; Brynjolfsson and Mitchell 2017)

In addition to using tools like the Suitability for Machine Learning checklist, running AI pilot projects and growing these projects in an iterative fashion in close collaboration with domain experts is another way to manage expectations. As a recent study of Deloitte and Tom Davenport revealed it is often more promising to start with "low-hanging fruit" projects that incrementally improve business processes than to directly go for disruptive "moon shots" (Davenport and Ronanki 2018). One interviewee from a global pharmaceutical company described this collaborative and incremental AI journey as follows: *"You know, there are problems [machine learning] can solve, and there are problems it cannot solve. So I think being clear with that from the start, and how [the project] will be an iterative learning process also for [business] [..] is very important. [..] [T]hey can't just say, 'Build me an algorithm that does this' and then go away"*.

Another approach preventing organizations from costly project failures is to cooperate with experienced AI or data analytics vendors. Such firms do not need to be large and well-known consultancies, but rather small and specialized firms that have a high industry knowledge. We talked to data scientists and managers of such vendors that are active, for example, in the energy retailing, publishing, or engineering industry. Clients of these vendors use them to outsource AI tasks that do not belong to their core business (e.g., text generation), to purchase industry standard solutions (e.g., benchmarks, customer valuation models), or learn from the vendor experience and use them as sparring partners.

1.3.2 Challenge 2: Managing AI Projects like Traditional IT Projects

When it comes to the development phase, many companies we talked to struggled with finding an appropriate project management methodology for running their AI projects. In contrast to traditional IT projects, which focus on building and deploying systems on time, on budget, and on scope, data science and AI projects are typically much more open and explorative. As one of our interviewees explained, this way of working is often new to business: *"business units know how classical IT projects are run: the IT department gets the requirements and then iteratively implements them. But data science projects require much more interaction between domain experts and data scientists and it is important that the business knows and understands this."* Due to their highly iterative and collaborative nature, many companies use stripped-down versions of agile develop-

ment practices like SCRUM to manage their data science and AI projects. However, these methods still assume that in the end a product will be delivered to the client. But data science and AI projects, in contrast, often deliver intangible outcomes. They are much more like research projects, involving framing the right questions, developing hypotheses, finding and extracting the right data, and running experiments to derive new knowledge or support decisions (Marchand and Peppard 2013).

The fuzzy front-end of data science projects, that is, defining questions and hypotheses, seems to be the most critical phase, as an interviewee from an international bank told us: *"I think the most important thing, and sometimes it is the most challenging, is to exactly define the issue business is facing and the outcome they are expecting. [..] Because if you don't define the expected outcome very clearly, you can go completely in the wrong direction."* Setting the course and monitoring progress in such open-ended projects is challenging and often creates tensions between data scientists and management: *"the biggest challenge are the people who want intermediate insights. They [always] want to know what you're up to and what you're finding [..]. That would be the managers and project leaders. What we then sometimes do, which is this actually bad practice, is we work for a whole week on solving the real problems and then the last day before the weekly catch-up meeting, we do something that they want to see"* (Data Scientist from global pharmaceutical company). The interviewee even went so far to say that *"the hardest part in this space is not doing the work [developing the ML algorithm and model] ... It is the connection between the data scientists and the project managers, that's very frustrating for me"*. To overcome the communication problems between data scientists and management, two obvious, but not that easy to implement, enablers exist. First, many companies try to improve their data scientists' communication skills. For example, many of the more technical interviewees in our study mentioned the importance of being a good (data-driven) storyteller and being able to provide convincing narratives for complex technical issues. Second, companies also started training their project managers in machine learning techniques. Not necessarily to be able to create machine learning models themselves, but to gain a better understanding about the differences to traditional software development.

Successful data science projects also do not end with the go-live of a system, but when the system generates new insights or employees use it to make data-driven decisions. These are things that are hard to monitor and, hence, *"it's necessary to spend a lot of time on defining how [..] to measure success"*, one interviewee emphasized. An example for good measurement comes from an insurance company we interviewed. They defined three interlinked but separate criteria for the success of its new machine learning-based churn prediction system (we illustrate them in Table 1). At the data scientist level, the predictive accuracy of the system was the key metric. At the business unit level, the development of actual churn rates in response to the decisions and actions proposed by the new system was the main success criteria. And finally at the executive management level, the predicted monetary customer lifetime value of the customers who could be prevented from churning was used as a key performance indicator of project success.

Stakeholder	**Key question**	**Metric**
Data Scientist	How accurately can I predict whether a customer will churn in the next 3 months?	Predictive accuracy of a machine learning model for churn prediction
Business Unit	How can we prevent customers from leaving us?	Development of churn rates after deployment of the system
Executive Management	How can we maximize the lifetime value of our customers?	Predicted net profit from all future transactions with customers who could be prevented from churning

Table 1: Example of a set of interrelated metrics for measuring the success of a churn prediction system at an insurance company

1.3.3 Challenge 3: Data Availability and Quality

The currently dominant approach for building AI systems is based on supervised machine learning algorithms. In this approach, machines learn from examples in a similar fashion like children do. That means that large sets of training data in the form of input-output pairs (e.g.,

credit applications and information on whether the customer was able to pay back the loan or not, machine usage data and information on subsequent errors or breakdowns, medical images and the corresponding diagnosis) are used to teach a system to correctly predict the output given a specific combination of input values. Current machine learning algorithms typically need tens of thousands or sometimes even millions of such training examples in order to work reliably. Deep neural networks are especially data hungry, as one data scientist of a company from the construction industry pointed out: *"Because we are producing so many different products, we do not have a lot of independent data samples [for each product]. We simply do not have enough data to do some of the more brute-force neural networks."*

It's not surprising that extracting or collecting the right data from the environment represents a major bottleneck in most AI projects. An interviewee from an international truck manufacturer summarizes what probably most practitioners in the field would confirm: *"Data quality is of course always a topic. No matter where I worked over the last years, it was always the case that 70 to 80 percent of the time was spent for data preparation and preprocessing."* While almost all interviewees agreed that data availability and quality is one of the most difficult challenges in AI projects, the concrete forms of data quality problems and their root causes seem to vary from company to company and project to project. For example, projects that tried to leverage data generated by physical machines to build, for instance, predictive maintenance systems were often confronted with a complete lack of training data. Heavy machines and equipment like ship engines, production lines, or buildings typically last several decades and are simply too old to collect and store data in digital form. A possible strategy for generating training data in such cases is to retrofit a couple of pilot machines with Internet of Things (IoT) capabilities such as sensors and wireless network access. However, this often means that one has to wait a considerable amount of time, sometimes years, before a sufficiently large and representative amount of training data has been collected. In other cases, the task to be automated was not supported end-to-end by a single IT system, or the performed steps were not logged in the required level of detail, and, therefore, it was difficult to collect the required training data. A strategy to obtain training data in such situations is the "learning apprentice" approach (Mitchell et al. 1990). Here, the AI sys-

tem acts as an apprentice watching the human experts performing their tasks and recording all relevant input and output data. After observing several thousands of repetitions of the same task, ideally performed by different individuals, the system is then able to learn the function required to correctly transform the input data into the output data; often better than the human experts did before. However, like in the retrofitting situation, it may take years before enough training data has been collected. Other projects relied heavily on data from Enterprise Resource Planning (ERP) systems and reported that—besides technical problems with extracting, transforming, and loading the data from the source systems—they had problems interpreting the correct meaning of existing data. For example, many sales departments create sales orders in their ERP systems to make ad-hoc stock reservations for customers. Once the customer has ordered the material, or does not need it anymore, they cancel these sales orders again. Including such "shadow" orders in training sets for recommender systems or price optimization models can easily lead to biases in machine learning models. The problem here is not that the required training data does not exist or has missing or wrong values, but that the data does not represent what it seems to represent. And, as one interviewee explained, the root cause of this problem is that sales employees *"did not predict that someone else would use the data afterwards"*. An enabler to overcome this challenge is to start creating a company-wide data culture that increases all employees' awareness of the importance of data quality and the willingness to share and reuse data across departments. Finally, companies who primarily interact with their customers via digital channels like websites or mobile apps reported the least data availability or quality problems. In many cases, these channels were already created with the purpose of data collection in mind.

1.3.4 Challenge 4: Interpretability

The latest generation of machine learning algorithms, especially deep neural networks, possess remarkable predictive power. They are, for example, able to detect malign mutations of cells on histopathological images, predict whether customers are about to switch from one online service to another based on transaction histories and clickstreams, or translate texts between different languages. However, they also have their limitations and drawbacks. One of the most

significant challenges when it comes to extracting knowledge from data and translating this knowledge into decisions is the lack of transparency behind the logic of how these models map inputs into outputs. Complex neural networks are black boxes often containing tens of millions of parameters that jointly define the function for translating inputs into the desired outputs. It is impossible for data scientists or end users to comprehend and interpret how these models make predictions. In other words, many AI systems have superhuman predictive capabilities, but are unable to explain the why and how behind their predictions. However, as one of our interviewees explained, *"[f]or some industries, it's really necessary to understand how the decision was made or the insight was brought to the user. Before [my current job] I was working with image analysis for the medical industry. And this is one that requires really a good understanding, because the doctor will not recommend something without understanding how that decision was made, or at least trusting that the decision was done in the right way"*. Similarly, a data scientist from an insurance company stated that *"we cannot provide a black box prediction model to our insurance brokers. A simple binary classification of which customers are likely to churn, and which are not, is not sufficient for them. Brokers need to know which data point has which influence on the likelihood of churn"*. This opinion was echoed by many of our interviewees across industries; when it comes to supporting knowledge workers in their decision-making processes through AI, *"traceability is the most important criterion"*. In some countries and industries there already exists a legal right to be given an explanation for the outputs of algorithms. For example, The European Union General Data Protection Regulation states that for automated decisions that significantly affect an individual (e.g., online credit application, e-recruiting practices) the subject of the decision should have the right "to obtain an explanation of the decision reached" (European Union 2018). In the United States, similar rights exist in the context of credit scoring.

Broadly speaking, there are two alternative technical enablers to overcome the interpretability challenge (Du et al. 2019). First, instead of using black box deep learning models, one can use less complex models, like rule-based systems or statistical learning models (e.g. linear regression, decision trees). These systems are intrinsically interpretable, even by people without a PhD degree in computer science. However, the increased interpretability of these models comes

at the cost of sacrificing some predictive accuracy. Second, one can develop a second model that tries to provide explanations for an existing black box model (for an example see Figure 5). This strategy combines the predictive accuracy of modern machine learning algorithms with the interpretability of statistical models. However, developing so-called post-hoc interpretability techniques that possess high explanation fidelity and end-user friendliness is still an emerging research area.

Figure 5: Example of the outputs of a post-hoc interpretability model providing an explanation for why a convolutional neural network has classified two animals on a picture as a dog (relevant pixels highlighted in red) and cat (relevant pixels highlighted in green; Ribeiro et al. 2016).

1.3.5 Challenge 5: Causality

Even if one can interpret the relationship between inputs and outputs of a machine learning model, there is no guarantee that the inputs are actually the causes of the output. Consider the following example from one of the cases we analyzed in our study. A digital marketing agency used deep neural networks to predict the success of Instagram posts that involve product placement by Internet celebrities. They found that product placement in front of mountain scenery is associated with higher numbers of likes and comments by fans. Should they recommend all their clients to shoot pictures with mountains in the background? Probably not. It just

happened that some of their most famous clients specialized in marketing sports and fashion products for the winter season. So the real cause for the success of their posts was their huge base of fans and followers, and not the background of the images they posted. This example nicely demonstrate that, when it comes to making decisions informed by AI systems and translating these decisions into real-life actions, it is indispensable to know whether an identified pattern is a real cause-and-effect relationship or just a spurious correlation, as intervening on variables which are spuriously correlated with an intended outcome will have no effect. The ride-hailing company Uber serves a good example for how firms can combine data science techniques with classical behavioral science methods to overcome the causality challenge (Totte Harinen and Bonnie Li 2019). Whenever possible, Uber conducts randomized control trials to investigate whether hypotheses generated through data science methods actually hold in real life. However, in many situations obtaining experimental data through A/B testing is simply not possible. For example, when introducing a product or service innovation, it is often not possible to compare its effect to an appropriate control group. In such situations, one has no other option than to work with existing observational data. One way to investigate causal questions with observational data is to model the causal process that has generated the data as thoroughly as possible and use appropriate statistical methods and tools to control for confounding factors (Pearl and Mackenzie 2018). Discovering the data-generating causal graph and investigating the nature and strength of its relationships requires close collaboration between data scientists and domain experts, who have substantive knowledge of the business problem in question. Figure 6 shows a possible causal data-generating graph for the Influence marketing example.

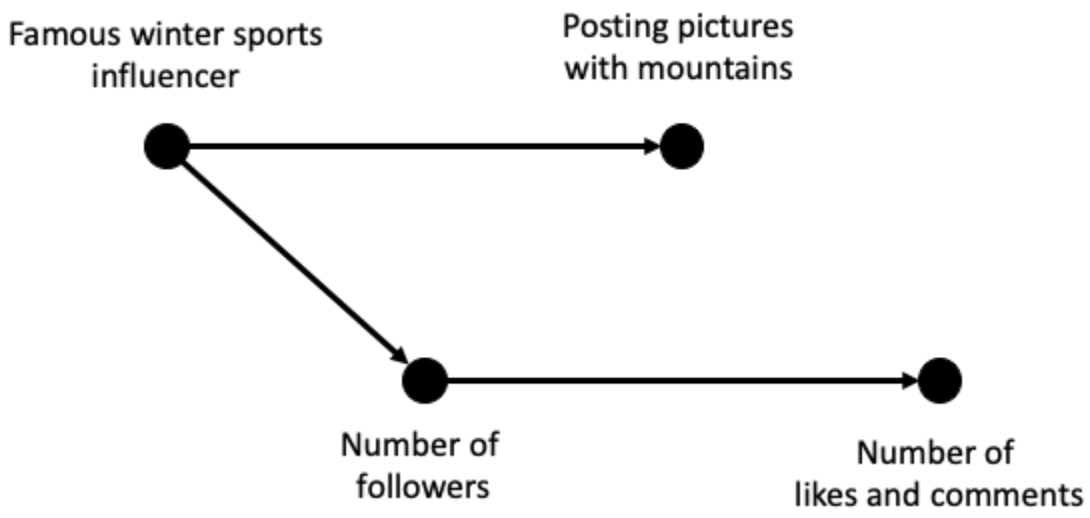

Figure 6: Example of a possible data-generating causal graph showing the causal relationships between variables in the Influencer marketing example. Being a famous winter sports influencer is causally related to posting pictures with mountains and having a large number of followers. Furthermore, a larger number of followers is causally related to the number of likes and comments a posted picture receives. If one does not control for the number of followers and being a famous winter sports influencer, it seems as if a mountain on a picture causes the number of likes and comments to rise.

1.3.6 Challenge 6: Missing Link to Transactional Systems and Processes

The second most often cited challenge in our study was, after lacking data availability and quality, the difficulty of feeding decisions made by AI systems back into a company's transactional systems and business processes. It almost seems that if there is an insurmountable barrier between the systems supporting or automating decision making and the systems that actually take actions. One of our interviewees phrased it like this: *"Our ERP system is my biggest headache in the whole world. It takes six months to get data out of it [..] and, yeah, we can't put data back into it."* In many cases, the problem of communicating between newly developed AI systems and legacy operational systems is simply a matter of missing interfaces. Large international firms are often running their business processes with enterprise systems that are 10 or 20 years old and have been developed on a completely different technology stack. A workaround to enable AI systems to better interact with transactional back-end systems that lack open interfaces is Robotic Process Automation (RPA). An RPA system watches the user execute a business process in the graphical user interface (GUI) of, for example, an ERP system, and is then able to automate the

process by repeating the tasks performed by the user directly in the GUI. The RPA system could, for example, copy & paste the outputs of an AI system into an ERP system, while simultaneously checking simple rules and customizing the process execution depending on predictions made by the AI system.

But the cause for the notorious missing link between analytical and transactional systems is not always a technical one; sometimes it is rooted in business or legal regulations. In one case we investigated, the data science department of a global pharmaceutical company developed a new machine learning model to dynamically choose the optimal packaging for outgoing shipments. However, the data scientists later discovered that in the pharmaceutical industry one needs yearly government approval for what packaging materials can be shipped to which countries. So the packaging recommendations generated by the new system were never used in the operational business process and the project was stopped.

In other cases, the reasons for not adopting the recommendations of AI systems are more of an organizational or even individual nature. In both the insurance and utilities sectors we discovered cases where sales representatives received reports with customer churn scores, indicating how likely it is that specific customers will cancel their contracts in the near future. Objectively, these scores should represent highly valuable input for the daily work of the sales reps. However, these scores seemed to be rarely used, either because the company's business processes were not designed to take such information into account (e.g., other conflicting business rules for prioritizing and contacting customers exist or it is not intended to proactively contact customers at all) or because the recipients were not data literate enough to interpret the information. One enabler to overcome this challenge is to find other ways for presenting the outputs of AI systems to users. In the churn case, for example, an electricity provider prescribed the questions their call center agents have to ask customers based on the customer's predicted likelihood of churn. So instead of showing call center agents a statistic, they gave them actionable instructions that are almost impossible to work around.

1.3.7 Challenge 7: Dynamic Environments

As described earlier, most of the current AI systems are based on supervised machine learning models. These models learn a function that is able to correctly map inputs into outputs; for example, a function to output the correct English translation for a given French input text, or a function that can predict the popularity (output) of an Instagram fashion post (input). Once the function has been learned with sufficient accuracy, it can be used to automate tasks that have traditionally been performed by human experts such as interpreters or marketing managers. This approach assumes, however, that the function to predict outputs based on inputs is stable over time (Gama et al. 2014). In the example of translating French to English this assumption is largely met. In the social media marketing example, however, the factors that determine a popular post change at least as fast as fashion and popular culture trends change. In other words, what used to be hip on Instagram six months ago, may not trigger half as many "likes" today. Sensing such structural changes and assessing their consequences for predictions and decisions made by AI-driven systems is a challenge that is often overlooked by companies, especially those that are new to machine learning. A data scientist from an international jewelry manufacturer and retailers we interviewed put it like this: *"You don't put something into production and then just keep it running. It's very much about continuous monitoring and figuring out that there's a drift in the data. [..] Having in-house consultants or data scientists in house is quite important here, because you need this continuous monitoring. Often [external] consultants come in and they do the project and deliver it and then they go somewhere else. [..] Traditional software keeps functioning in the same way over time. But the performance of machine learning models might degrade when the data changes."* The statement highlights two things. First, the importance of continuously monitoring the performance of predictive models. This is often more difficult than it sounds, as it requires to have access to up-to-date ground-truth data which can be used as a benchmark for the algorithmic predictions. Second, the importance of clearly defined organizational roles and responsibilities. Simply buying a standard software package or hiring external consultants to build a predictive model is not sufficient. Successfully employing an AI system is not a one-time project, but a continuous effort. Even a company like Google learned this the hard way, when their famous Google Flu Trend service, which was able to predict the spread of the flu based on what user type into Google's search box, started to wildly overpredict flu levels (Lazer et al. 2014).

One of the main reasons was that Google's engineers never updated the underlying machine learning models, although both the nature of influenza pandemics and the logic and usage of its search engine changed considerably over the years.

A holistic approach to overcome this challenge is the concept of MLOps (Machine Learning Operations; Talagala 2018). Inspired by practices from software development to reduce the time between committing a change to a system and the change being successfully placed into normal production (DevOps), MLOps aims at improving the quality of machine learning models in production by, on the one hand, increasing automation in monitoring and improving the performance of these models, and, on the other hand, implementing organizational structures for ensuring compliance with business and regulatory requirements. Concrete practices include, for example, the use of drift detection techniques to notice variations in the input data, conducting A/B tests to detect degradation or side-effects in the real-world impact of models, or interdisciplinary teams spanning technical and business experts to ensure quality and minimize the risk of AI systems.

AI Value Creation Process	Challenge	Examples	Enablers
Planning	Inflated management expectations	Executive management thinks every knowledge work task can be automated by AI; Management directly wants to go for "moonshot" projects	Check task's suitability for machine learning, Engage analytics vendors, AI pilot projects
Development	Managing AI projects like traditional IT projects	Project managers expect a continuous delivery of results; Data scientists waste time creating intermediate products and reports for management	Data scientists with good communication skills, Data-literate project managers, Appropriate KPIs
From Envi-	Data availability	No training data available for	Collecting training data

ronment to Data	and quality	creating a supervised machine learning model; Data scientists have to spent the majority of their time on extracting and cleaning data from source systems	from the field, Learning apprentice approach, Company-wide data culture
From Data to Knowledge	Interpretability	AI system can predict future events, but not explain how the predictions were made; Only the machines learn, not the humans	Intrinsically interpretable algorithms, Post-hoc explainability methods
From Knowledge to Decisions	Causality	Identified patterns are just correlations and no cause-and-effect relationships; real-world interventions have not the predicted effect	Randomized control trials, Causal modeling
From Decision to Actions	Missing link to transactional systems and processes	For technical or regulatory reasons the decisions suggested by an AI system cannot be fed back into transactional systems; Employees ignore outputs of AI systems	Robotic Process Automation, Data scientists with awareness about compliance regulations, Prescriptions instead of predictions
Operations	Dynamic environments	Predictive accuracy of machine learning models degrades over time; In-house employees do not know how to update or retrain models	MLOps

Table 2: Overview of Challenges and Enablers along the AI Value Creation Process

Exhibit: How energy retailers pursue the path towards AI value creation

Energy suppliers have been existing for over a century and serve a vital market for everyone. By their very nature, they have access to a great number of customers, are even monopolists in

some areas. Digitization does not stop at these incumbents, rather affects utilities with all its might: They have to build data-heavy smart grids and collect vast amounts of data as a result of digitizing meter-to-cash and other business processes. We spoke with five AI professionals about eleven projects in the energy retailing industry and found both failure and success stories:

With their grown and strong hierarchies, we observed a discrepancy between management expectations and the frontline business (*Challenge 1*). Ambitious AI projects that are imposed by top management are prone to collapse, as AI technology cannot be easily bundled into a standard software product with clear interfaces, like energy management systems, ERP, or reporting tools, which can be purchased and just set up (*Challenge 2*). Another issue is that operational processes in marketing, customer contact, and other functions are not ready to act upon insights from AI prediction models (*Challenge 6*). There are also often strong concerns about privacy and regulatory issues that inhibit bottom-up data-driven innovation (*Challenge 7*). In addition, several projects fail because the necessary data is not available or not in the shape to be used (*Challenge 3*). The aims to predict rare events, e.g., outage of energy installations, or to analyze phenomena that have not been recorded in the past, e.g., customer preferences or contract cancelation reasons, are just two examples of such data issues.

Nonetheless, we saw many good examples of how these companies create value with AI. Several energy retailers cooperate with specialized analytics vendors to buy-in sophisticated solutions, for example, for churn monitoring, customer valuation, or smart meter data analyses instead of developing them from scratch. Using externals as sparring partners, they build their own AI competencies and wisely select which task to in- and outsource (*Enabler: Engage analytics vendors*). Also, management exercises—with vendors or in the form of pilot projects—to properly evaluate AI initiatives in order to make wise decisions (*Enabler: AI pilot projects*). Other firms cooperate with universities and kickstart risky analytics projects with students. The successful companies, of course, dedicate appropriate—IT and non-IT—resources to provide the necessary infrastructure and data access for these projects (*Enabler: Data-literate project managers*). AI applications are often computationally lightweight procedures in the background rather than new IT systems that need heavy maintenance and user training (*Enabler: MLOps*). A com-

mon feature in all successful projects was indeed an entrepreneur in the company who was pushing the data-driven and AI-related projects (*Enablers: Company-wide data culture*).

Several issues regarding the use of AI technology in business demand further investigation. One issue is the continuous model development and retraining. When, for example, prediction models of customer behavior are used to improve advertising campaigns, the predictive quality decreases over time as the 'top' customers are served (*Challenge 7, Enabler: Randomized control trials*). Another issue is the application of models that are trained to predict but are used to explain phenomena in practice, for example, in driver or root-cause analyses (*Challenge 5, Enabler: Causal modeling*).

Concluding comments

Since the inception of the computing field, organizations have tried to use IT for automating various tasks and business processes. Highly structured and repetitive tasks like payroll processing or order handling have been automated a long time ago using systems that rely on hand-coded rules and knowledge bases. Yet, due to the fragile nature and the enormous effort required for creating and maintaining rules and knowledge, these first-wave AI systems were not suitable for automating non-routine tasks. With the increased diffusion of second-wave AI systems that are able to inductively learn how to perform a task without being explicitly programmed, more and more knowledge work tasks can be taken over by machines. However, this new generation of AI systems require different implementation and management approaches. Successfully implementing AI in an organizational setting is not a one-time activity and requires more than deploying IT systems. AI systems need to be closely interlinked with operational systems, their predictive performance needs to be continuously monitored, their decisions need to be checked with regards to interpretability and causality, and they have to be retrained and adapted when their environment changes.

Appendix

#	Type of company	Size of company	Experience of IP	Gender of IP	Project description
1	Media	small	1	M	Automatic writing
2	Heating manufacturer	medium	4	M	Call center automation / classification (App for staffing)
3	Private insurance	medium	>7	M	Automated insurance sales
4	Fashion retailing	small	3	F	Customer target groups for email campaign
5	Jewellery retail and manufacturer	medium	3	M	Personas modeling and activation
6	Data analytics vendor	small	3	M	Churn consulting with AI model
7	Data analytics vendor	small	7	M	Market segmentation and persona generation
8	Truck industry	large	2	M	Text analysis on component replacements
9	Truck industry	large	2	M	Complex visualization
10	Media	medium	>7	F	Management support

11	Media	medium	>7	F	Automatic media summary
12	IT consulting	large	5	F	Automatic processing of health care policies and summarizing the content
13	Media	small	1	M	Publisher consulting
14	Data analytics vendor	small	3	M	Churn project in Germany with regular reports and integration in call center system
15	Heating manufacturer	medium	4	M	Predictive maintenance (identify backup heating system is in place)
16	Robotics	small	3	M	Robotics training data collection
17	Robotics	small	3	M	Analytics vendor projects
18	Mechanical engineering	small	7	M	Machines for metalworking company
19	Mechanical engineering	small	7	M	Optimizing pneumatic machines
20	Data analytics vendor	small	7	M	Churn score integration in call center system
21	Private insurance	medium	>7	M	Risk prediction

22	Private insurance	medium	>7	M	BI Analyses
23	Original equipment manufacturer	medium	>7	M	Spare parts demand forecasts supply chain
24	Original equipment manufacturer	medium	>7	M	Causal impact analysis for Supply Chain
25	Energy retailing	medium	1	M	Rule identification for email classification
26	Energy retailing	medium	1	M	Price forecast
27	Energy retailing	medium	1	M	Meter defect detection
28	Energy retailing	medium	1	M	Public transportation app
29	Insurance and pension	medium	2	M	Risk prediction health insurance
30	Transportation	medium	2	M	Travel journey analysis to inform decision makers about routes
31	Pharmaceutical	large	1	F	Demand forecasting supply chain
32	IT consulting freelancer	small	>5	F	Logistics data warehouse integration
33	Insulation manu-	medium	3	M	Factory Optimization

	facturer				
34	Jewellery retail and manufacturer	medium	3	M	Classification of database reports to categories (mapping) with NLP
35	Electronics manufacturer	large	4	M	Public transport demand forecast
36	Electronics manufacturer	large	4	M	Parking area demand forecast
37	Consulting	small	2	M	Employee termination risk (using social network analysis)
38	Consulting	small	2	M	Employee satisfaction analytics (using social network analysis, smart watch data)
39	Media	medium	4	F	Identifying article metrics for journalists
40	Media	medium	4	F	Recommendation engine for journalists with application
41	Media	medium	4	F	Recommender system for news articles
42	Pharmaceutical	large	2	M	Customer communication channel scoring

43	Banking	medium	?	F	Debt collection (scores from two prediction models, fixed list of customers to call)
44	Bank	medium	3	F	Fraud detection (prediction model plus mathematical function, case worker takes action on)
45	Life insurance	medium	1	M	Churn prediction
46	Media	medium	2	M	Influencer search engine
47	Jewellery retail and manufacturer	medium	3	M	Marketing impact evaluation of measures (A/B-Tests, …)
48	Original equipment manufacturer	medium	>7	M	Predictive maintenance (aftersales)
49	Original equipment manufacturer	medium	>7	M	Predictive maintenance (service agreement)
50	Online retailing	large	3	M	Truck scheduling for ordering system
51	Online retailing	large	3	M	Root cause analysis for delivery shortage
52	Food waste saving	medium	5	M	Customer Journey

53	Energy retailing	medium	5	M	Market analysis battery storage
54	Energy retailing	medium	5	M	Quantitiy structure for electricity purchase
55	Energy retailing	medium	5	M	Chatbot for customer interactions

Table 3: List of projects and details of interview partners (IP)

2 Paper II

Shifting AI Value Creation Mechanisms: An Explorative Study

By Arisa Shollo,

Konstantin Hopf,

Tiemo Thiess,

and

Oliver Müller

Abstract

Advancements in artificial intelligence (AI) technologies are rapidly changing the competitive landscape. In the search for an appropriate strategic response, firms are currently engaging in a large variety of AI initiatives. However, recent studies suggest that companies are falling short in creating tangible business value through AI. As the current scientific body of knowledge lacks comprehensive frameworks for explaining this phenomenon, we conducted an empirical analysis of 57 applications of AI in 29 different companies. We identified three broad types and five subtypes of effective AI value creation mechanisms that represent unique and fundamental sources of value. Furthermore, we found that organizations dynamically shift from one value creation mechanism to another by reconfiguring their AI applications. We also identified necessary but not sufficient conditions for successfully leveraging the different AI value creation mechanisms that provide an alternative explanation for the current high failure rate of AI projects.

Keywords: Artificial intelligence, value creation mechanisms, knowledge creation, augmentation, automation, AI strategy, interview study

2.1 Introduction

Artificial intelligence (AI), broadly understood as the capability of machines to perform cognitive tasks at a level that is comparable or even superior to humans (see, e.g., Russell and Norvig 2009), is a rapidly advancing general-purpose technology that holds the potential to reshape the nature of work (Frank et al. 2019). Recent years have seen breakthroughs in the field of AI in terms of basic research and development. However, despite the rapid technological advances, AI's impact on organizations and economies has been modest so far (Brynjolfsson et al. 2017). In fact, while researchers were able to measure positive productivity impacts of AI-related technologies, skills, and practices for companies of some industries (Müller et al. 2018), in some geographies (Tambe 2014), or for some types of business processes (Wu et al. 2019), the economic productivity growth of the overall economy has declined over the past decade (Brynjolfsson et al. 2017). According to Brynjolfsson and colleagues a likely explanation for this paradox are implementation and restructuring time lags. The full effects of AI will not be realized until companies develop and implement new complementary organizational capabilities in order to fully leverage the potential of AI; and the development of these complementary innovations takes considerable time.

Indeed, recent studies reported that organizations are struggling with realizing and sustaining value from AI initiatives (Brynjolfsson et al. 2017; Tarafdar et al. 2019). Most companies run only ad-hoc pilots or apply AI in just a single business process (Fountaine et al. 2019). One of the key enablers for creating and sustaining value from AI is moving from pilots to company-wide programs addressing end-to-end processes (Fountaine et al. 2019). However, anecdotal evidence suggests that many companies instead go for "moon shoot" projects without systematically evaluating automation needs (Davenport and Ronanki 2018).

Both, the academic and practitioner-oriented discourses are characterized by a strong focus on the opportunities AI provides for organizations, but neglect to uncover the mechanisms that

organizations need to realize promised value. Regarding the question of how organizations realize the potential value promised by AI, the discussion—primarily based on conceptual analysis—centers around the choice between AI augmentation and automation strategies that companies can follow, else known as the support-versus-replace debate (see Markus 2017; Zuboff 1985). However, the information systems field is currently lacking empirical research that analyzes what strategies organizations create to realize value from AI technologies (Günther et al. 2017), how they decide about the level of augmentation/ automation (Coombs et al. 2020), what kind of AI value organizations pursue, and what kind of value they actually get. As Markus (2017) succinctly puts it, information systems scholars should "move the debate beyond generic positions and provide nuanced advice" (p. 234) about the strategic management of algorithmic intelligence. Hence, the information systems field is in dire need of further empirical studies that carefully examine how organizations actually realize value from AI in practice.

Against this background, we address the calls from both information systems (Coombs et al. 2020; Galliers et al. 2017; Markus 2017; Rai et al. 2019) and management scholars (von Krogh 2018; Raisch and Krakowski 2020) to study the role of AI in value creation by investigating the following research question:

How do organizations create value through applications of AI?

To answer this question, we carried out an exploratory interview study focusing on applications of AI in organizations. We studied 57 of such applications in 29 organizations. To generate rich data about these AI applications, and how they contribute to value creation, we interviewed 40 data scientists, as they have been described as professionals with an interdisciplinary background and holistic view on AI and its application in organizations (van der Aalst 2014; Davenport and Patil 2012; Plastino and Purdy 2018).

Our findings show that there are different value creation mechanisms that organizations can employ to realize impact from AI. In particular, first, we identified three broad types and five subtypes of AI value creation mechanisms, which represent unique and fundamental sources of value organizations pursue. Second, we found that AI value creation is a dynamic process and,

in order to sustain value from applications of AI, organizations shift from one value creation mechanism to another by reconfiguring their AI applications. Third, we identified necessary but not sufficient conditions for successfully leveraging the different value creation mechanisms that provide an alternative explanation for the current high failure rate of AI projects.

The remainder of this paper is structured as follows: First, we provide an overview of the current discourse on AI value creation. Next, we present our research method, followed by the presentation and discussion of our findings. We conclude by summarizing the study's limitations and contributions.

2.2 Theoretical background

The use of AI in organizations has begun to attract significant attention from management and information systems researchers, who especially focus on the ways how organizations create value through AI (Brynjolfsson and Mcafee 2017; Coombs et al. 2020; Davenport and Ronanki 2018). Several studies focus on the automation affordances of AI technologies (e.g., Coombs et al. 2020; Tarafdar et al. 2019). However, the strong focus on automation through AI seems to fall short in terms of business productivity (Brynjolfsson et al., 2017) and social implications for the workforce (Brynjolfsson and Mitchell 2017; Galliers et al. 2017; Manyika et al. 2017; Newell and Marabelli 2015). Recent conceptual works describe different facets of embedding AI in organizations, be it platforms of human-AI hybrids (Rai et al., 2019), the ability of a new generation of information systems that can learn and act autonomously and thus form a new agency (Ågerfalk 2020), and the emergence of metahuman systems (Lyytinen et al. 2020). These conceptual works point to a knowledge gap regarding the best strategy for embedding AI in organizations and for deriving and sustaining value from it.

The discourse on the effective combination of human-AI configurations recently turned back to a key finding of earlier human-computer interaction studies (Parasuraman et al. 2000; Sheridan and Verplank 1978; Wickens et al. 2010), namely that AI value creation is not a binary choice between using AI for augmentation or for automation of human work. Rather, a nuanced view on both extremes is necessary. Raisch and Krakowski (2020) advocate that organizations need to

engage in both applications of AI that augment and automate human capabilities in order to create and sustain AI value in the long term. Similarly, Jarrahi (2019) discusses positive affordances and side effects of using AI for informating and automating work and concludes that "AI can ... play a key role not just by performing organizational tasks more efficiently but also by empowering workers through symbiotic interactions with humans … Workers must be given the opportunity to dynamically participate in the analysis and interpretation of AI-generated results" (p. 183). Shrestha et al. (Shrestha et al. 2019) illustrate the various possible human-AI configurations in organizational decision making processes, and Grønsund & Aanestad (2020) underline the importance of humans in the loop for auditing and altering practices when working with AI. However, extant research lacks clarity about how organizations should pursue the way towards the right portfolio of applications of AI. This problem is further reinforced by the fact that so many nuances of human-AI combinations exist (Faraj et al. 2018) which create new kinds of metahuman systems (Lyytinen et al. 2020).

In the range between, on the one hand, full process automation and, on the other hand, the support of human decisions (augmentation), prior work identified different types and applications of AI value creation. Davenport and Ronanki (2018, p. 110), for example, find that "automating business processes, gaining insight through data analysis, and engaging with customers and employees" are three types of AI value creation. AI can be applied for different augmentation purposes. One obvious way is to train machine learning (ML) models to provide predictions for operative business processes (e.g., churn scores at the individual customer level) and support the work of people by presenting them with these predictions. This is what we would intuitively refer to as true augmentation. The other way is to extract more general rules from models (e.g., the factors that drive customer churn), rather than individual predictions, and inform human work by presenting these rules to workers. This latter type is more knowledge generation than a support of the operational process, given that AI is used to create new knowledge. Exemplary forms of knowledge generation using AI are, for example, in research and development for generating new scientific theory (Berente et al. 2019) or for gaining insights in social science using explainable AI (Miller, 2019). The duality of AI technology as being

usable for automation and knowledge creation continues a longstanding perception of "the two faces of intelligent technology" (Zuboff 1985, p. 5), namely automating and informating. AI-driven value creation is often associated with automating (Coombs et al., 2020; Tarafdar et al., 2019), which might be explained by a logic that Zuboff already explained in the 1980s: "[Managers] narrow emphasis on automation is the web of economic logic in which they must operate. Conventional accounting formulas treat technology as capital substitution of labor. As many managers have learned, 'to justify a new computer we have to show job eliminations.'" (p. 12) This perception resulted in alarming reports about potential job losses through AI (Frey & Osborne, 2017), but AI value creation is more complex, as organizations need to continuously realign work practices, organizational models, and stakeholder interests in order to realize value (Günther et al. 2017; Lebovitz 2019; Markus 2017).

Although research on AI has received increasing attention in the past few years, research on how AI creates business value remains scarce. In their comprehensive literature review on intelligent automation, Coombs et al. (2020) point to the "significant limitations in our understanding of how organisations decide the level of Intelligent Automation" (p. 12). There is also an ongoing debate on how to assess the value of big data analytics (an AI-related technology) for firms, particularly in shedding light on how big data analytics investments can yield tangible business value (Grover et al. 2018; Günther et al. 2017; Sharma et al. 2014). This stream of research offers important insight on data driven value creation processes by proposing frameworks and research agendas. In particular, Grover et al. (2018) propose a value creation framework that links big data analytics efforts with necessary capability building and strategic value creation.

On the same subject, Günther et al. (2017) and Markus (2017) conclude that current literature on value creation from datification initiatives in firms is rather conceptual and that additional empirical research is needed to derive new theory and actionable knowledge for strategic decision making. In particular, to understand AI value creation in organizations requires analyzing how organizations "translate, as well as fail to translate, its potential into actual social and economic value" (Günther et al., 2017, p. 192), what kind of value organizations gain from applications of

AI in practice, and how they sustain value over time. In addition, to provide responsible practical recommendations about what organizations should do, we need to analyze the organizational and environmental conditions that promote or inhibit AI value realization (Markus, 2017).

2.3 Method

In response to the research question, we chose an exploratory research design. Specifically, we conducted a semi-structured interview study to gather rich in-depth insights. Due to the exploratory nature of this study, the research employs a qualitative approach, aimed at understanding the mechanisms of how organizations create value through AI.

2.3.1 Data Collection

We collected our data in two rounds. In the first round (between October 2018 and January 2020), we interviewed 40 data scientists about their involvement in 57 corporate applications of AI (on-going projects and deployed systems, a full list of the analyzed projects is included in the appendix Table A.1 and the interviews are listed in Table A.2). The interviewed data scientists performed different roles and, while they were sometimes specialized in different aspects of AI value creation (e.g. data engineering, data modeling, model deployment), overall, they were responsible for the implementation of machine learning algorithms. In this way, we were able to gain an inside, bottom up view into AI value creation (Plastino and Purdy, 2018) that we believe is lacking currently from the information systems literature. The retrospective accounts of the interviewed data scientists were the main data source for our data analysis and theorizing process. Following Gioia (2013), we treated these "informants" as "knowledgeable agents", giving them an "extraordinary voice" (p. 26). Practically, this means that we assumed that our informants "know what they are trying to do and can explain their thoughts, intentions, and actions" (Gioia et al. 2013, p. 17). In particular, we asked them about the AI projects that they were involved in, thereby grounding the interview in participants' own experiences. This helped to keep the interviews rooted in actual events and settings, which reduces the risk of the discourse spiraling into abstractions, generalities, and cultural scripts (Schultze and Avital 2011). The interviews followed an interview guide that we flexibly adjusted over time driven by the accounts

of the informants (Gioia et al. 2013) and the ongoing coding and theorizing process that we carried out with the help of a coding logbook. The overall structure of the interview guide was, however, stable and largely based around a conceptualization of an AI value creation process that proved to be useful for structuring the interviews. This process consisted of four stages, namely transforming "data to knowledge", "knowledge to decisions", "decisions to actions", and "actions to value" (Sharma et al. 2014; Thiess and Müller 2018). In particular, we asked the informants how they approached and experienced each sub-stage of this process, what difficulties they faced in each stage, and which strategies they used to create what they perceived as value created by AI. In some cases, we reached back out to informants of earlier interviews to ask about concepts that were arising from later interviews (Gioia et al. 2013). Our technical background and past experience in AI technology allowed us to go into the necessary depth of the domain with the interview partners, so that we understood how the projects were embedded in the organisational context and which AI technologies were used. Still, the interviews remained open to probing into informants responses when they initiated a new area of inquiry. The recorded interviews have a mean duration of 48.5 minutes (12 minutes standard deviation).

A second round of data collection (between January and June 2020) focused on the validation of our findings. For this, we organized a workshop with data scientists on January 16 (2020) with 13 participants (seven of them were interview partners, six were new participants). In this focus group discussion, we presented the nascent state of our theoretical conceptualizations. In addition, we asked earlier informants that could not participate for written feedback on our interim findings and received substantive feedback from ten informants.

We selected the interviewees purposefully, ensuring a high variety of industries (e.g., retailing, mechanical engineering, energy retailing, banking and finance, robotics, IT consulting, data analytics vendors, transportation, media, pharma, foods and beverages, public sector) and covering organizations of different sizes from small (0-15 employees) over medium (16-1.000 employees) to large companies (>1.000 employees). Interview partners were 21% female and had between several months to up to seven years of job experience.

2.3.2 Data Analysis

We recorded all interviews and transcribed them verbatim, which resulted in more than 600 pages of original text. Due to the exploratory nature of our research question, we applied an open coding approach to identify relevant and interesting chunks of data. During the initial rounds of coding, we tried to stay close to the words and phrases of the informants, which led to a large amount of open codes (Gioia et al. 2013). While the first round of open coding was largely data-driven, we used our experience and knowledge of relevant theory in the later rounds of coding to make sense of the emerging concepts. Thus, we condensed them into broader themes and overarching theoretical dimensions (Gioia et al. 2013), as exemplified in Figure 1. Following an inductive approach, we tried not to force the emerging concepts into frames of pre-existing theory. A more detailed selection of indicative quotes and codes are listed in the appendix in Table A.3 and A.4.

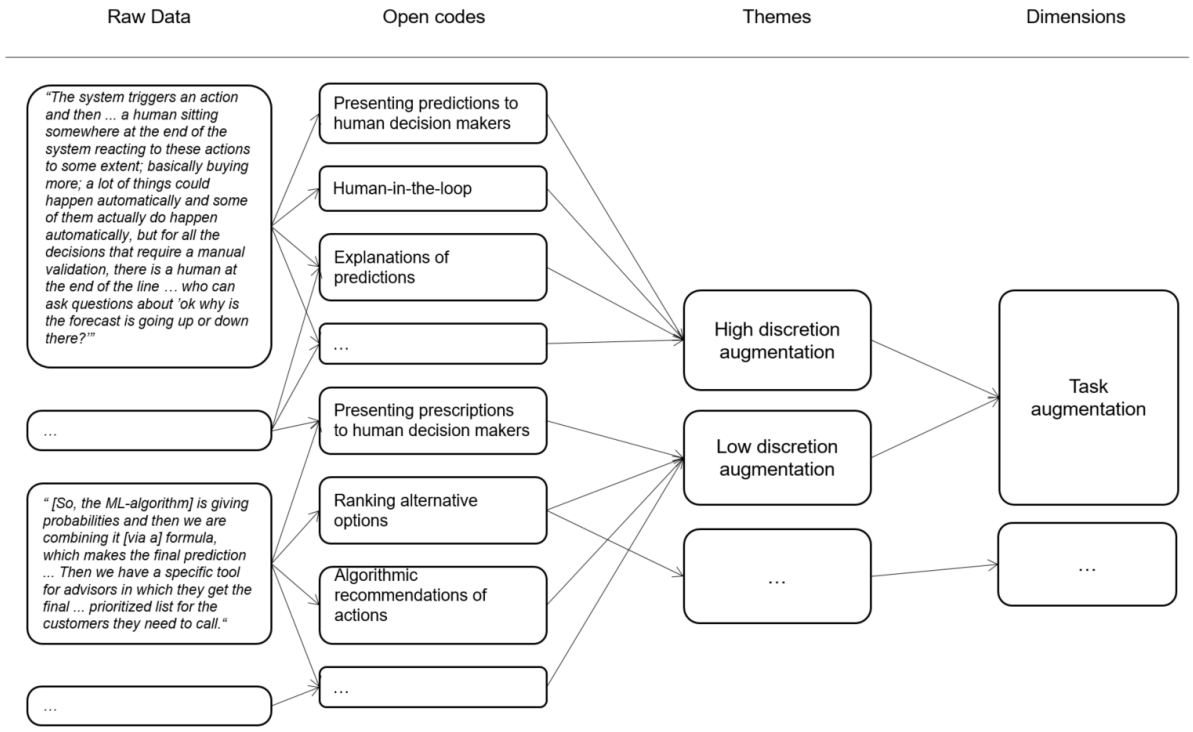

Figure 1: From raw data over open codes to themes and dimensions

Through an iterative analysis, themes and dimensions in relation to different strategies of creating value with AI, their characteristics, and use emerged as "transparently observable" (Eisenhardt 1989, p. 537). The first rounds of analysis resulted in three unique AI value creation

mechanisms, which we refined in further rounds of analysis and coding until we reached a point of theoretical saturation, in which no conceptual deviations remained (Glaser and Strauss 1967). Frameworks and concepts from existing literature supported this iterative data analysis process, in particular, we used the big data analytics value creation framework by Grover et al. (2018) as an analytical lens and vocabulary for our theorizing process. Once the emerging AI value creation mechanisms were sufficiently stable, we re-analyzed the data again through the lens of the emerging theory that is, we classified all projects according to their value creation mechanisms (see Table A.1).

This led us to uncover that over their lifecycle AI applications often did not only follow one mechanism of value creation, but sometimes organizations reconfigured their AI applications in order to shift to a different value creation mechanism. Such reconfigurations eventually became the focus of another round of data analysis, in which we searched for factors (conditions) that have caused the reconfiguration.

We used MAXQDA software to analyze all collected data (interview and focus group meeting transcripts, written feedback). Original quotes were translated, if necessary.

2.4 Findings

This section describes the results of our empirical analysis. We identified three distinct AI value creation mechanisms that we ordered according to the influence that AI has on decision making and process execution (first low, then moderate to high, then very high). Thereafter, we exemplified how companies sometimes reconfigure their AI applications and, thereby, move between different value creation mechanisms. In order to explain these reconfigurations, we identified necessary conditions for shifting towards another AI value creation mechanism, as we describe in the final subsection. We observed that a mismatch between a value creation mechanism and its necessary conditions is the main reason for project failure and that managers, who understand this, reconfigure their AI applications to change the intended level of augmentation/automation and strive for a different type of value.

2.4.1 AI value creation mechanisms

We present our findings following the concept of value creation mechanisms proposed by Grover et al. (2018) to express how the companies we studied created value from the application of AI. In the process of translating tangible and intangible IT assets (e.g., hardware, software, human resources) to measurable business value (e.g., performance improvement in terms of increased productivity), value creation mechanisms are—according to Grover et al. (2018)—a mediator between capabilities enabled by assets (e.g., the ability to collect and manage data and train predictive models from data) and value targets (e.g., better decision making). These value creation mechanisms represent fundamental sources of value being pursued by a company. Specifically, we found three AI value creation mechanisms with in total five subtypes. We illustrate them in Figure 2 and describe them below.

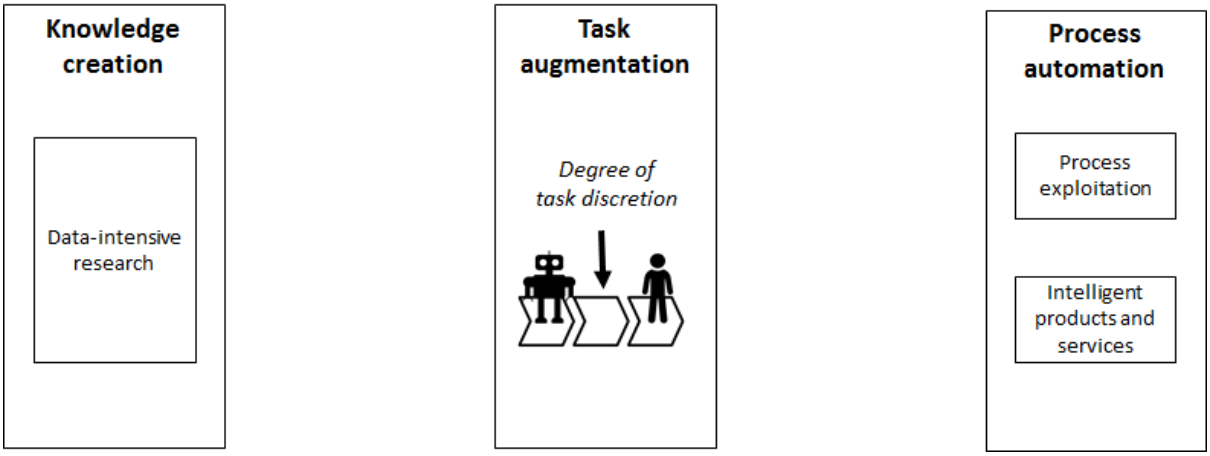

Figure 2: AI value creation mechanisms

2.4.1.1 Knowledge creation (AI value creation mechanism 1)

Although AI is often associated with the notion of automation, a significant number of AI applications in our sample (26 of 57) did not focus on automation. Their goal was rather to apply AI for creating new knowledge, by utilizing machine learning algorithms that inductively identify trends and patterns in the data and learn rules on how to predict outcomes based on a given set of input variables. In this way, AI machines are able to learn from data, identify the rules themselves instead of being programmed by humans and continuously improve over time.

We found that these knowledge creation projects often supported rather strategic decision making processes (e.g., data-driven segmentation for market research, root-cause analysis for delivery shortage). We further found that these applications of AI rarely ended directly in the implementation of productive IT systems and instead focused on easy-to-implement prototypes, or simple reports that have a relatively low risk of failure. However, project execution was often inefficient due to the ad-hoc nature of these projects and the fact that the analysis outputs were not used in a clearly defined downstream decision making task, but rather used as inputs for strategic decisions.

AI for knowledge creation often happened in the form of data-intensive research projects: one-time, often ad-hoc, analyses of historical data to generate exploratory and explanatory knowledge by identifying rules for mostly strategic decisions. These projects used diverse data sources and applied a variety of data analysis methods to collect data (including IoT infrastructures, field experiments, web scraping) and analyze them. Data analysis methods ranged from traditional statistical models, such as linear or logistic regression, to supervised and unsupervised ML algorithms, but also included time series forecasting and simulations. Overall, we could observe a tendency to apply transparent algorithms like logistic regression instead of applying highly complex black boxed algorithms like neural networks, as the goal of these projects was to explore relationships between variables or to test existing hypotheses with data. Many interviewees described the objective of these projects as "identifying the drivers of the process", which then allow to generate an understanding of existing phenomena and support the development of new concepts and ideas. One of the interviewees reviewed from his work:

"In my opinion, what generates the most impact are the factors influencing customer behavior that our models spit out. They serve as a basis for discussion in the company about how one might develop new concepts." (Translated; P03: Customer churn modeling)

Data-intensive research projects typically create knowledge that explains phenomena for whole (sub-)populations (e.g., customer groups) rather than predicting events on an individual level (e.g., a single customer). Such knowledge is usually represented as estimated parameters of a

statistical model that, for example, explains transaction fraud or customer churn. In one of the analyzed projects, for instance, a rule was identified that for a one-unit increase in customer spending, the likelihood that a customer churns decreases by a certain percentage point. In most cases, such analyses are highly contextualized to a specific setting and only executed once, or just a few times. The data collection and analysis often follow approaches that are commonly employed in science, namely econometric modeling of observational data or experiments (A/B tests). A data scientist at a global jewelry retailer described how they use A/B testing to obtain data on the behavior of customers:

"So in terms of revenue, we do uplift testing … basically, we run a business as usual [scenario] versus our new approach. So splitting ... our budgets ..., half is going [to the new approach] and then half of the budget [is going to the] traditional approach ... then holding those two outcomes up against each other and taking the delta." (P46: Marketing impact evaluation)

2.4.1.2 Task augmentation (AI value creation mechanism 2)

We found that task augmentation projects are mostly concerned with implementing ML-based systems into production environments to support everyday downstream tasks. Here, humans remain the decision makers in the end, while the implementation of the decisions can be mandated to either human or machine actuators (see middle part of Figure 3). Task augmentation projects in our sample (18 of 57) created value by using ML-based systems to exploit scarce resources (e.g., employees) in a more optimal manner by augmenting the limited human information processing capabilities with the superhuman information processing capabilities of AI systems. The projects were mostly about augmenting tasks that are concerned with programmable decision making processes on a tactical or operational level. Moreover, we found that ML-based systems for task augmentation most commonly output individual-level predictions (e.g., scores or forecasts) and prescriptions (e.g., recommendations) based on supervised ML and optimization algorithms.

We identified two subtypes of the task augmentation mechanism: applications that leave low or high discretion for the human decision maker (i.e., the user of the system).

2.4.1.2.1 High-discretion augmentation

These applications of AI support decision making in ways that leave large parts of the final judgment and choices to the discretion of their users. This is done, for example, by displaying a set of predictions or recommendations, from which the user can select one or decide to ignore them completely. The sales department of a pharmaceutical company, for example, has enriched its customer relationship management (CRM) system with color-coded ratings that indicate the preferred communication channel for office-based physicians. Based on these predictions, but also on the experience and personal relationships of the sales agents, they can choose the most promising channel for contacting the client. Another example of high-discretion task augmentation is a process mining system deployed in a factory of a large building materials manufacturer. The system alerts production experts with push-notifications about certain steps in production processes that do not operate as planned, thus helping them to detect process inefficiencies early. The system, however, tells neither why the irregularities occur nor does it give explicit prescriptions about what to do, this remains at the discretion of the production experts:

"The system .. provides some support for the process experts ... So basically instead of having no idea where to look, we gave them [a] shortlist which will say this based on the data, 'it could be in this area or this area' ... And then they can keep diving into ... it." (P31: Factory optimization)

2.4.1.2.2 Low-discretion augmentation

These applications of AI allow users only little influence on the final decisions made. For example, an energy retailer implemented a task augmentation system in its call center, which dynamically instructs agents what questions to ask and in which order. The agents have little influence on how to conduct the conversation, as they continuously receive prescriptions by the system on how to act and have only a few seconds to think about questions and answers. In the case of such low-discretion augmentation systems, the main role of the user is not to be a decision maker, but rather to be an actuator that can, indeed, intervene and overrule the system when exceptions or anomalies occur. We found that tasks that are supported by such low-discretion systems could be automated rather easily in theory, because users mostly follow the instruc-

tions of the system, but in practice, those processes remained at least partially manual, due to technical, ethical, legal, or risk-related concerns. Another example of a low-discretion augmentation system that we investigated was a forecasting system deployed for inventory management at a large original equipment manufacturer. The system recommended how many units of a particular material its users should purchase. Here users could only accept or reject the system generated recommendations:

"The system triggers an action and then ... [there is] a human sitting somewhere ... reacting to these actions to some extent basically replenishing more; A lot of things could happen automatically and some of them actually do happen automatically, but for all the decisions that require a manual validation, there is a human … that can ask questions about 'ok why is the forecast is going up or down there'?" (P21: Spare parts demand forecasts supply chain)

In contrast to high-discretion augmentation systems, which usually make predictions, low-discretion augmentation systems go one step further and generate prescriptions. They do so through decision functions that allow automatic selection of a final and often mathematically optimal set of courses of actions out of several evaluated alternatives. Usually, optimization algorithms (e.g., linear programming) or simple heuristics like "take the alternative with the highest score" are applied. For example, in the call center system mentioned earlier, the task augmentation system only instructs agents to ask questions about customer satisfaction at the end of a call, if the estimated churn probability for a given customer exceeds a certain threshold. Therefore, the call center agent has almost no discretion in this case. Similarly, a large bank uses decision functions to combine predictions from two ML models with information about customer debts:

"we are getting two probabilities. We have a specific formula in which we are combining these two ... we get the final number, and we rank that number. So the customers with the highest values, they are our top priority … [this] makes the final decision ... Then we have a specific tool for advisors in which they get the final ... prioritized list for the customers they need to call." (P41: Debt collection)

2.4.1.3 Process automation (AI value creation mechanism 3)

Some applications of AI in our sample (13 of 57) aim for process automation. The objectives of the underlying projects were to develop, operate, and maintain ML-based automation systems. Such systems usually apply supervised ML algorithms to execute continuously end-to-end business processes without human intervention. Moreover, we saw that they, just like low-discretion augmentation systems, rely on decision functions to choose a finite and often mathematically optimal set of courses of actions. However, we observed that, in contrast to knowledge creation and augmentation systems, process automation systems implement actions directly without a human in the loop (see Figure 3).

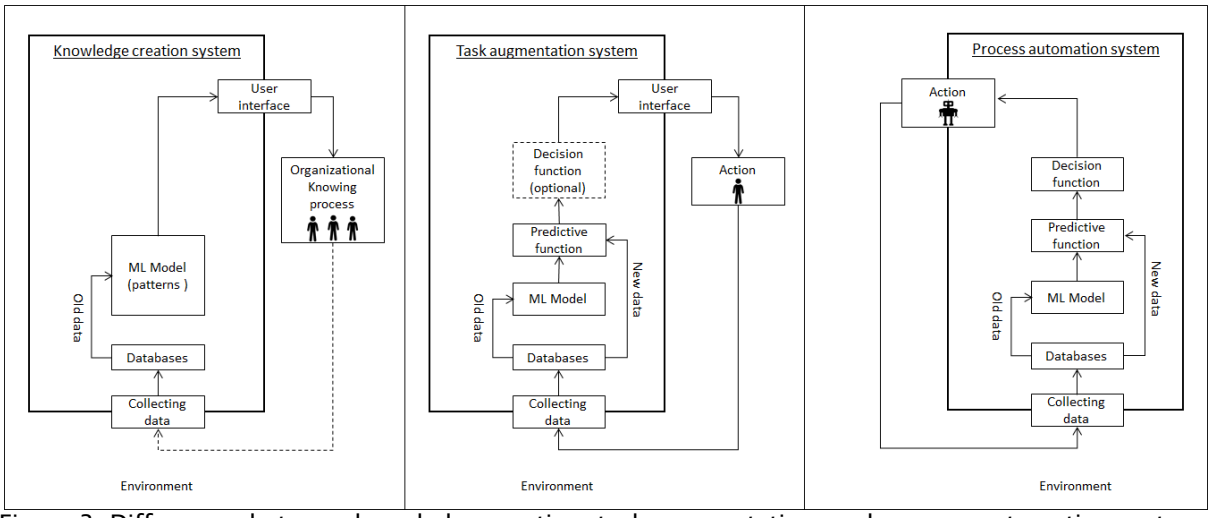

Figure 3: Differences between knowledge creation, task augmentation, and process automation systems

Given that automation systems often replace human labor, one can demonstrate the value of such systems by comparing the efficiency of the automated process with the priory deployed human-centered process. A data scientist at a media service provider (P08: Automatic media summary) summarizes: "*We use AI for automation and I think it's easier to see the value ... if you look at the insights and strategies, it is not that easy.*" She further exemplifies: *"we have the minimum work ... per profile. [Employees] have a specific time to spend in each profile. So, in the minute these profiles will go live, now fully automated, they can estimate the value in terms of reduction of cost per customer."*

We identified two subtypes of AI based process automation systems. AI can improve, on the one hand, the efficiency of a business process by reducing time and resources needed for process execution. We name these applications of AI *process exploitation*. On the other hand, AI can

be an integral part of an intelligent—often called smart—product or service. We label these applications of AI as *intelligent products and services*.

2.4.1.3.1 Process exploitation

With automation through AI, intelligent systems take over complete parts of internal business processes and make them more efficient. A media service provider, for example, whose business it is to automatically generate media summaries for clients (so that they can see what the press wrote about them), used ML to automate one of their core business processes. They developed *"an engine that can classify an article as relevant or not for a specific [customer] profile"* (P08: Media summary). This activity, which was up until now done manually, was completely replaced by the automation system. Another example is an application of AI in which a startup company tries to automatically generate penny novels with deep neural networks instead of human authors:

"One and a half, two years [ago], the first ... neural networks were able to [generate] more or less meaningful texts ..., the four founders said 'okay, ... we are bringing in AI to generate texts and sell these texts.' ... I would have to lie if I said our in-house systems are already writing novels. But the daily goal we have is ... to put it bluntly, [to write] these penny novels." (Translated; P11: Automatic writing)

Data scientists developing systems for AI process exploitation have to circumvent many challenges which requires them to think creatively and to experiment with many parameters of neural networks:

"*How do we manage to ensure that this [neural] network is able to pursue a storyline in the long term, also to always mention the same character, the same protagonists and that this also holds true, is consistent, and coherent? ... [Do we need] to learn the structure of a book through a meta-model and then fill this structure with life afterwards? Or does it make sense to generate individual chapters separately, which are then put together by another network? ... We are currently able to influence the number of protagonists a bit. This parameter [influences] how creative this network should be. … if you give it too much freedom, it usually ends in a doomsday dystopia."* (Translated; P11: Automatic writing)

2.4.1.3.2 Intelligent Products and Services

These AI applications create new intelligent products and services that can be offered directly to customers. The offers may be sold (e.g. smart devices, apps, or paid services) or designed to increase the service level of existing offers (e.g. to improve customer service). As such, they must be well-embedded into existing offerings. Typical to these novel offerings is that they would not be feasible without AI technology. An example of this are chatbots that process customer inquiries, as a data scientist at a energy retailer explains:

"*We wanted to settle certain standard inquiries via a chatbot in Facebook Messenger or via WhatsApp ... so we internally tested several chatbots from various providers ... in the end we did a benchmarking which chatbot best understands what is happening and then answering this customer request without the customer calling us*" (Translated, P55: Chatbot for customer interactions).

As we observe, the quality of service must be sufficient for these intelligent products and services to be deployed. Sufficiency, of course, depends on the criticality of the service or the product function e.g. a chatbot does not necessarily need to be precise all the time in contrast to a demand prediction that might have more severe consequences for the organization.

We summarize the overall characteristics of the three identified value creation mechanisms in Table 1 and differentiate them according to attributes that we described before.

Table 1: Attributes of the AI value creation mechanisms

	Knowledge creation (AI value creation mechanism 1)	**Task augmentation (AI value creation mechanism 2)**	**Process automation (AI value creation mechanism 3)**
Sub-types	Data-intensive research	Augmentation of low- and high-discretion tasks	Process exploitation, intelligent products and services
Value targets	Organizational knowing	Better decision making	Increased productivity, novel value offerings

ML Output	Explanations (patterns, rules, relationships)	Predictions (scores, forecasts etc.) and prescriptions (recommendations)	Prescriptions (recommendations)
Types of ML	Hypothesis-testing, causal inference, simulation, clustering, A/B testing	Supervised ML and optimization	Supervised ML and optimization
Decision maker	Human	Human (at least in the loop)	Machine
Action taker	Human	Human or machine	Machine
Type of decision making	Non-programmed	Programmed	Programmed
Level of decision making	Strategic and tactical	Tactical and operational	Operational

2.4.2 Reconfiguring the AI value creation mechanisms

Our data indicate that the value creation mechanism of an AI application can change over time. Over-ambitious AI automation projects, for instance, might already fail in the pilot project stage. Data-intensive research projects, on the other hand, can evolve into AI systems that augment human tasks or fully automate processes. In the following, we describe three selected AI initiatives whose value creation mechanisms have been reconfigured repeatedly over time (we schematically illustrate their path in Figure 4).

2.4.2.1 Observed reconfigurations

We present three different examples that give a rich account of how organizations reconfigured value creation mechanisms below. We found that the dominant trajectory is from left to right (knowledge creation to automation), but there are also examples that move in the opposite direction.

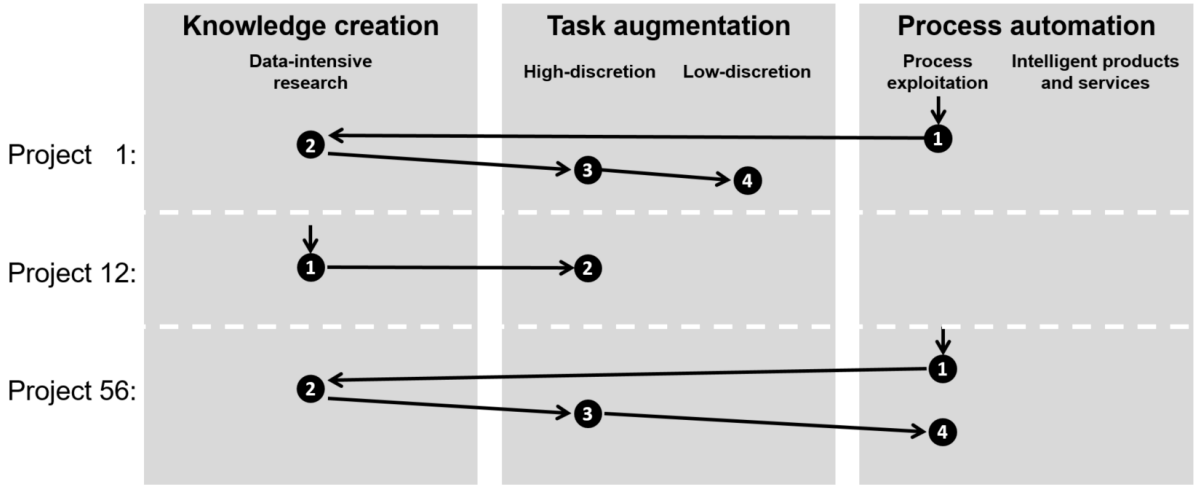

Figure 4: Selected reconfigurations

The first reconfiguration example stems from energy retailing. Here, a third-party analytics vendor developed an AI-based process automation application for retailers to identify customers that are likely to cancel their contract. Based on the system's predictions, customers with high churn probability should be automatically approached with targeted advertising. The churn predictions worked successfully in the lab using historical training and test data, but it turned out that the system had no effect on actual customer churn rates in the first field tests. In addition, the utility company had technical problems feeding the predictions back into their CRM system, which sends out advertising mails to customers. As a reaction to these difficulties, the AI project was reconfigured from a process automation system to a data-intensive research project that aimed at identifying the drivers of customer churn. Once such drivers were identified, a pilot was built and field tested in order to investigate how customers would react to targeted advertising in a churn context. After several rounds of development and testing of the new predictive ML model, both the predictive accuracy of the model and the estimated business impact of its implementation increased steadily. Eventually, the company integrated the churn scores into their CRM system via a newly built interface. After that, the scores could be made accessible in the company's call center system to augment the tasks of sales professionals by prescribing them to ask certain questions, if the churn score of a customer exceeded a certain predefined threshold (high-discretion task augmentation).

The second reconfiguration example is a project from a heating manufacturer that had to optimize their call center after facing a social media shitstorm where heating installers complained about long waiting times. The CIO instructed a team of IT professionals and a data scientist to improve the call center operations. In a data-intensive research project, the team used process mining to analyze the call center logs and found several wrong call routing configurations. As a side-project, they tested how well the volume of calls for certain product lines can be predicted over time in order to improve the staffing and vacation planning, given that maintaining heating installations is a seasonal business. This analysis eventually resulted in a ML-based application for call center scheduling and staffing (high-discretion task augmentation system).

The third example stems from a global jewelry retailer that had already successfully automated its online advertisement process (process exploitation). However, during the Covid-19 pandemic, it found itself in new uncharted waters. The predictions of their ad placement algorithms were suddenly not as accurate anymore as they used to be, leading to unsatisfactory, if not disastrous, advertising performance. After the first week of the pandemic, the company suspended their automatic advertising and all marketing-related decisions were taken based on human judgment. Meanwhile, the company started a data-intensive research project by collecting sales data about countries, which were hit early by the pandemic (i.e., China, Italy). Based on these data the company created interactive dashboards that would allow the data scientists to gain quickly an understanding of the changing customer needs and behaviors:

"We have taken all our online and offline sales traffic, and created a timeline and looked 'during what events did sales start to change?'. Also, we have looked at all our media efforts: 'When did they leave?'. Then we created a picture based on it, to find the right model." (P56: Online advertisement process)

Based on these data and insights, they then started to train new predictive models tailored to the pandemic situation. In the third week of the pandemic, the company created a new decision making team—they called it the SWAT team—acting across the owned and paid media channels of the global organization using the output of the new predictive models as information for their decision making. Hence, while AI was applied to create knowledge about customer needs

(data-intensive research) in the beginning, it transformed to being used for routine decision making (high-discretion task augmentation) over time, i.e., high-discretion taks augmentation. The initiative was still ongoing, but the mid-term goal was to return to the old mode of AI-based process automation (process exploitation).

2.4.2.2 Conditions

After identifying the above-described reconfigurations, we analyzed the data again, searching for factors that explain why such reconfigurations occur. As a result, we found that the reconfigurations can be explained by a set of necessary, but not sufficient, conditions that need to be fulfilled in order to successfully leverage the identified AI value creation mechanisms. These conditions can be internal (e.g., necessary assets or capabilities) or external to the company (e.g., necessary environmental factors). The conditions are illustrated in Figure 5; they are inherited from mechanism to mechanism from left to right (i.e., the knowledge creation mechanism has two necessary conditions, the task augmentation mechanism has five necessary conditions, the process automation mechanism has eight necessary conditions).

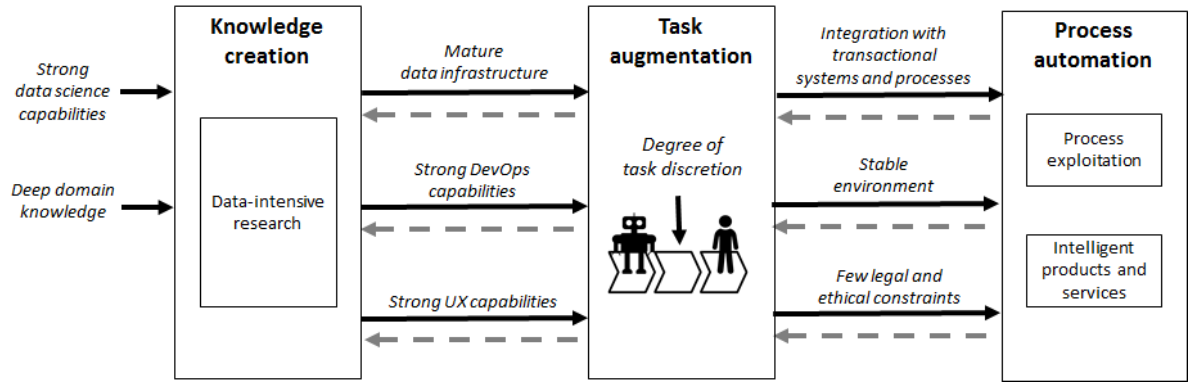

Figure 5: AI value creation mechanisms *and their necessary but not sufficient conditions*

We identified two basic conditions that are necessary to realize value from any of the three AI value creation types, namely *strong data science capabilities* and *deep domain knowledge*.

Strong data science capabilities are, on the one hand, necessary for collecting and transforming the necessary data for AI applications and to allow for rigorous model development and evaluation. On the other hand, they are necessary to be able to communicate the often complex out-

comes of ML algorithms to a broader business audience: *"it's ... a task of a data scientist to be able to present the analysis and the results in a way that's meaningful and understandable to all relevant stakeholders"* (P57: Revenue forecasting in supply chain).

However, technical capabilities alone are not sufficient to succeed with AI projects. *Deep domain knowledge* is equally important, as it enables people to understand the data generating process—the true (underlying) phenomenon that is creating the data—that they are trying to model. Although modern ML algorithms allow for a largely data-driven approach to modeling, deep domain knowledge is the only way to detect causal mechanisms in observational data. It can both be brought in via data scientists' own domain knowledge or via domain experts from the business, which data scientists can consult. A data scientist at a large jewelry retailer emphasized:

"What is very challenging is to understand the data itself. Because my background is not really from retail, it's not really from marketing. So, I have to know at least 'what does this data mean?' Is it reasonable to have this kind of output or what kind of input can I use?" (P45: Customer target groups for email campaign)

Beyond these two basic conditions, we identified three necessary conditions to move from the knowledge creation mechanism to task augmentation mechanism, namely *mature data infrastructure*, *UX capabilities*, and *strong DevOps capabilities.*

We found that a *mature data infrastructure* is necessary to assure mid- and long-term success of task augmentation. Without a mature data infrastructure and the high quality data it provides, even the most advanced AI technology cannot be fully utilized. A data scientist working on personas modeling (target groups) describes this dilemma as follows:

"We have a Data Lake and I think technology-wise we're quite advanced. We spend a ton of money on technology, but in terms of maturity and onboarding of data and processes around data and data quality, super, super basic, because it [has] never really been a priority for now." (P47: Personas modeling).

While a strong data foundation is a key enabler of AI, our informants reported that a carefully designed *user experience (UX)*—the way how a user interacts with and experiences a system—

may be equally important. Especially for task augmentation applications, in which there is always a human in the loop, a positive user experience can avoid phenomena like algorithm aversion. One enabler for achieving a positive experience is to avoid the sole use of incomprehensible black box algorithms, either by using transparent models (e.g., linear or logistic regression, decision trees, and rule-based systems) which are inherently interpretable for users or by adding post-hoc explanation capabilities to highly complex model types like those based on neural networks. The following statement of a data scientist at a large bank exemplifies the need for interpretable machine learning algorithms in task augmentation systems:

"In most cases [users] ask like, 'Oh, but how does the algorithm work?' or 'Why did the algorithm give this low score or this high score?' ... That is one of our challenges, because as per se, most of the ML models are black boxes. You do not know exactly what happens there." (P42: Fraud detection)

Finally, we found that for sustaining long-term value realization of task augmentation systems, it is necessary to continuously monitor and improve data pipelines, models, interfaces, and actuators. Therefore, *strong development and operations (DevOps) capabilities* are necessary to assure the quality of ML-based systems. Such capabilities allow, for instance, raising warnings and adjusting systems whenever significant performance drops are observed.

"You don't put something into production and then just [run it]. It's very much about continuous monitoring and figuring out if there's a drift or changes [in the data]. … Traditional software keeps functioning the same way over time. Whereas ML models might degrade or other stuff happens to [it when] the data changes. I think there's more complexity than the underlying system [alone]." (P32: Document matching natural language processing)

For successfully implementing and running AI-based process automation applications, we identified three necessary conditions, namely *integration with transactional systems and processes*, *stable environment*, and *few ethical and legal constraints*.

We observed that many AI projects struggle with implementing automation systems, because they can simply not achieve a seamless *integration with transactional systems and processes*. While it is usually possible to read data from transactional systems in a more or less timely manner, it

can be difficult to write new data back into them. Even if the required legacy programming skills (e.g. Cobol, ABAP) were in place, it was hard for some projects to integrate AI systems with old transactional systems, simply because of lacking interfaces. Another reason that we found were governance structures that do not allow data science teams to make required changes to transactional systems.

For automation systems, it is essential that they are deployed in a sufficiently *stable environment*. One of the main issues that we observed was that ML models were built under the assumption that the data generating process is stationary. Changes to transactional systems, business processes, or the organizational environment, however, can change the underlying data generating process. For example, frequent price changes in business-to-consumer settings are known to change consumers' buying patterns and, hence, will have negative consequences for the predictive accuracy of existing demand forecasting models. Such performance losses due to unstable data generating processes can draw a whole project into question, as one data scientist reported that their system after months in production showed *"a drift in the performance. And then [the stakeholder] didn't trust us as much as after that."* (P32: Document matching natural language processing)

Lastly, we observed that the success of process automation systems could be dependent on the presence of *ethical and legal constraints*. While knowledge creation and augmentation systems always operate under human supervision (humans make final decisions and / or take actions), process automation systems do not have any humans in the loop when making procedural decisions. Hence, it is essential to design and monitor these systems carefully to ensure that they act within the ethical and legal boundaries. The CTO of an analytics vendor gave an example of a situation in which they did not want to automate a process fully due to legal and ethical concerns:

"A client said recently: 'It's totally cool if you can automate this, but we will never do it, we always want to have a manual step in it... because then it's not profiling … because of data protection, compliance ... otherwise every customer has to be informed and [has to] agree." (P01: Churn prediction)

In addition, at a large pharmaceutical manufacturer, a data scientist was experiencing difficulties to implement an already built process automation system due to legal constraints by the government: "*we built a model to optimize the shipment packaging, and that would be dynamic, so each shipment would come in, and it would tell it the most optimal packaging. But we have what's called standard operating procedures that [need to be] approved by the government [they are] approved yearly ..., so we're having trouble, once we build it, to implement it.*" (P29: Supply chain optimization)

2.5 Discussion

Inspired by recent studies that indicate that AI value creation remains an ambition rather than a reality for many organizations (Brynjolfsson et al., 2017; Davenport & Ronanki, 2018; Fountaine et al., 2019; Ransbotham et al., 2017; Tarafdar et al., 2019), we investigated the AI value creation processes. In particular, we analyzed 57 applications of AI in 29 different companies and unpacked how organizations implement and create value from AI. We found that AI cannot only be used to improve business processes or create new products and services (Tarafdar et al., 2019), but also as a means of generating new knowledge. In this respect, we identified three dominant AI value creation mechanisms, each providing a unique source of value. In particular, we found that organizations use AI a) to create new knowledge by discovering correlations, identifying possible causal relationships, and testing the predictability of events, b) to augment tasks through hybrid AI-Human collaborations and, c) to automate entire processes. For each of the three AI value creation mechanisms, we identified subtypes, which illustrate the high variety of human-AI assemblages. The five subtypes augment and automate organizational processes and practices in different ways and at different levels. In the following, we discuss our findings, first, from an organizational decision making perspective, and second, through the lens of organizational AI use, that has been at the center of recent scientific discourse.

2.5.1 AI support in organizational decision making

From an organizational decision making perspective, the first value creation mechanism (knowledge creation) primarily supports strategic decision making processes. Data-intensive research projects create business knowledge. Hence, they enable organizational knowing (knowledge creation and learning) through the application of AI (Lyytinen et al. 2020; Shollo

and Galliers 2016). The second AI value creation mechanism (task augmentation) exploits the knowledge creation, learning, and predictive capabilities of AI technologies in tactical and operational decisions. In particular, the low and high discretion task augmentation subtypes are both about supporting routine decision making tasks. The two types differ in terms of how the division of labor between humans and AI is organized. The task augmentation value creation type refers to the role that AI plays in hybrid sequential organizational decision making structures (Shrestha et al. 2019). The identification of the subtypes further highlights that the division of labor in these sequential structures need to be taken into consideration in order for value to be realized. The third mechanism of AI value creation (process automation) includes creating value by automating end-to-end internal processes and creating new intelligent products or services for customers. Here, operational decisions and actions are automated by AI machines either in a form of an end to end process (decisions of where and when to advertise in the World Wide Web) or in the form of an intelligent service or product (e.g., conversational agents as new user interfaces; Davenport and Ronanki 2018; Shrestha et al. 2019).

2.5.2 Nuances of AI augmentation and automation

The empirical findings of our study support the recent scholarly discourse on effective human-AI configurations by reinforcing the observation that AI value creation is not a binary choice between using AI for augmentation or automation (Parasuraman et al. 2000; Sheridan and Verplank 1978; Wickens et al. 2010). The identified mechanisms subtypes can be placed along a continuum. This continuum has, on the one end, a high degree of automation through AI. In the middle is augmentation, where the focus lies on humans and AI working together to make processes more efficient or decisions better. At the other end is knowledge creation, which is a low level of augmentation, where AI is used as a tool for (mostly) one-time analyses to find new patterns or working hypotheses. In this respect, we are able to locate the identified subtypes on the augmentation-automation continuum that is suggested by recent literature (Grønsund and Aanestad 2020; Raisch and Krakowski 2020; Shrestha et al. 2019) from data-intensive research (mostly augmentation) to intelligent products or service (mostly automation). Hence, with this

study, we further unpack the continuum by identifying some of the shades of how AI creates value in organizations.

Previous conceptualization of AI affordances in terms of augmentation and automation (Brynjolfsson and McAfee 2014; Daugherty and Wilson 2018; Davenport and Kirby 2016; Günther et al. 2017) imply an either/or strategic choice on how AI can be used in organizational settings. Our findings, however, provide evidence that the choice organizations face is not just one of augmentation and automation, but rather of the many different shades of the augmentation-automation continuum. Hence, for organizations to create and sustain value from AI applications, they need to choose carefully where to be and how to navigate this continuum (Parasuraman et al. 2000; Sheridan and Verplank 1978). Organizations need to engage in all three, knowledge creation, augmentation, and automation through AI in order to realize its potential in the long term (Raisch and Krakowski 2020). Treating AI-driven augmentation-automation as a continuum allows organizations to perceive and, hence, pursue multiple levels of augmentation and automation, their distinctive benefits, and leverage them separately rather than focusing on the augmentation and automation dichotomous categorization. Embracing the dichotomy while acknowledging the need for both AI applications, Raisch and Krakowski (2020) advocate for organizations to "purposefully iterate between distinct automation and augmentation tasks, allowing long-term engagement with both forces" (p. 19). Our findings confirm this purposeful iteration, albeit it is taking place between the different levels of the augmentation-automation continuum, that is, in a more granular level. In this way, it allows organizations short-term engagement with different levels of automation-augmentation. For example, depending on the complexity of the case or the model's predictive accuracy an AI automation system would instantly shift to operate as an AI augmentation system forwarding the case to a human decision maker. We illustrate the three AI value creation mechanisms and their subtypes on the AI value creation continuum in Table 2 and list also the streams of literature that support these findings, even if the literature streams were not necessarily linked so far.

Table 2: Identified AI value creation mechanisms with subtypes, illustrative examples, and references to related studies

AI value creation mechanism	Subtype	Illustrative examples	References
Knowledge creation	Data-intensive research project	Extract knowledge (hypotheses, novel patterns) from data	Berente et al. (2019)
↕	High-discretion augmentation	Supporting managerial decision making through predictions, algorithmic preselection of alternatives	Miller (2019)
Task augmentation	Low-discretion augmentation	Prescriptions through prediction models and optimizations	Davenport and Ronanki (2018), Shrestha et al. (2019)
	Process exploitation	Automated algorithmic decision making and autonomous action taking	Grønsund and Aanestad (2020)
↕	Product and service development	Novel "smart" services and products (e.g., voice assistants, chatbots)	Raisch and Krakowski (2020)
Task automation			

2.5.3 Necessary but not sufficient conditions as guidelines to navigate through the AI value creation process

For each value creation mechanism, we identified necessary but not sufficient conditions (see Figure 5 for an overview). As these conditions change continuously in organizations for a variety of different reasons (internal as well as external forces that operate in and upon the organization), we argue that organizations need to reconfigure continuously their AI initiatives for AI value to be realized. Our findings highlight different trajectories that applications of AI follow in their pursuit of value realization. This is a process with many potential paths from knowledge creation to process automation, but also the other way around. While the left to right trajectory (i.e., moving towards AI process automation) is more frequent and accepted as the "desired evolution", other trajectories are not met with enthusiasm. One reason is that pro-

ject re-scoping is avoided, as it can be interpreted as the inability of project managers or data scientists to handle the project (Hartl and Hess 2019). Conceptualizing AI value creation as a process (Markus and Robey 1988) moving between different AI value creation mechanisms based on a set of necessary but not sufficient conditions, we unpack organizational AI value creation processes. Here, we highlight "contextual contingencies" (Markus 2017). In particular, we provide an empirically based explanation of why and how companies engage in different levels of the augmentation-automation continuum as conditions change.

The identified reconfigurations emphasize the dynamic nature of AI value creation and provide an evolutionary perspective on "how organizations manage algorithms, not just in the planning levels or during implementation, but also after organizations have gained considerable experience in using algorithms" (Markus 2017, p. 236). In this respect, our findings support that strategic decisions about the AI value creation mechanisms are not static but have to change based on organizational and environmental conditions in order to sustain value from AI applications (Markus 2017).

Further, the finding that organizations need to configure and reconfigure their AI applications continuously to match changing conditions points to a dynamic capabilities perspective of AI application management (Božič and Dimovski 2019; Daniel et al. 2014). As systems that learn, AI machines are different from other IT applications, hence, organizations need to develop new dynamic capabilities to be able to leverage their potential while responding to a dynamic competitive environment. They are required to sense changes in the environment, assess their impact on the AI effectiveness, and accordingly make necessary changes. Especially crucial to monitor are data generating processes that produce training data for learning algorithms, because computational models will be altered in that way and make future predictions, prescriptions or even decisions. Future studies should focus on identifying first order and second order capabilities of firms, both from an AI portfolio perspective (Daniel et al. 2014) but also from an AI dynamic capability perspective (Li and Chan 2019) to assist AI managers and AI units in appropriating business value from AI resources by influencing a set of AI-related ordinary capabilities.

We concur with previous discussions of viewing AI machines as a new class of agents in organizations rather than mere artifacts (Ågerfalk 2020; Floridi and Sanders 2004; Raisch and Krakowski 2020). Our empirical conceptualization of AI value creation mechanisms contributes to this discussion by further exploring the kind of tasks and acts that they perform. This concretely means that by exploring the tasks that AI machines perform on behalf of humans, we can begin to understand better the far-reaching consequences of their agency (Ågerfalk 2020; Rai et al. 2019). Our findings are useful in implementing the research agendas in that we have empirically confirmed that the cooperation between humans and AI can take place in any configuration. People as well as AI can take the lead role or perform tasks completely alone. In no case, however, is this an either/or decision. While our value creation mechanisms remain agnostic when it comes to the (negative) consequences, they do shed light on the increased spectrum of acts that AI based systems can perform. Given this wide spectrum, what becomes critical is the question of who is accountable for these actions performed by AI based systems—a growing research area for information systems researchers and not only (Ågerfalk 2020; Rai et al. 2019).

2.5.4 Implications for Practice

Our research has two main implications for practitioners. First, the identified AI value creation mechanisms and conditions are an effective management tool for strategic decisions, for both, the formulation as well as the execution of an organization's AI strategy. Business and IT executives can use the AI value creation mechanisms to recognize the actual value contribution made by their AI projects. They can also clarify and communicate the positioning of current and planned AI projects and their contributions to organizational value. This can help to avoid situations where applications of AI are evaluated based on, for example, performance measures that do not apply to their particular types. In addition, we underlined that AI applications can have diverse value targets (reaching from knowledge creation, process efficiency gains, until the development of new offerings). We therefore encourage firms to set up a portfolio of AI initiatives in order to not only generate the maximum value from AI, but also to create the different types of value that are feasible. Our second implication is that our findings, especially the contextual conditions, inform executives, but also AI project managers, about possible problems

and mitigation approaches, and provide decision support on the AI value creation mechanisms adopted by projects as well as the development steps that need to be undertaken for sustaining AI value as conditions change. This is particularly important, as a change in conditions means that value realization is at risk. Hence, AI strategists and managers have to re-adjust to different mechanisms of AI value creation, for which the conditions for value realization can be met. Hence, our framework provides support to "organizations decid[ing] the level of Intelligent Automation " (Coombs et al. 2020, p. 12) of their AI applications by choosing the appropriate AI value creation mechanism. Another important insight for managers is that failing to realize AI process automation does not equate with losing value.

2.6 Conclusion

Over the last years, information systems and management scholars have repeatedly called for empirical research on the strategies that organizations employ to realize value from AI (e.g., Coombs et al. 2020; Galliers et al. 2017; Markus 2017; Rai et al. 2019; Raisch and Krakowski, 2020; von Krogh 2018) that goes beyond "generic positions and provide[s] nuanced advice" (Markus 2017, p. 234). With this study, we contribute to research on the organizational use and business value of AI by proposing a theoretical framework of value creation mechanisms, re-configurations, and conditions of AI applications that is grounded in rich empirical data and prior literature. Our study provides fine-grained empirical evidence on the different levels of the augmentation-automation continuum as well as on how and why organizations shift between the different levels. In summary, our findings assist in understanding why many organizations struggle with implementing AI and extracting tangible and sustainable business value from it. At the same time, it provides practical guidance for managers to navigate the largely uncharted waters of AI value creation.

In this study we go as far as identifying AI value creation mechanisms. However, due to the explorative nature of this research, the identified AI value creation mechanisms are possibly not exhaustive. Further, we cannot quantify the exact value that AI creates in these organizations. Explicitly, we do not capture the functional and symbolic impact these AI applications have on organizational performance (Grover et al. 2018). Future studies, similar to those on the impact

of big data analytics on firm performance (Müller et al. 2018; Tambe 2014; Wu et al. 2019), should add an economic point of view on the impact of AI applications in organizations.

Reflecting on our methodology, and as AI applications are being increasingly diffused in organizations, we acknowledge that interviewing more stakeholders involved in AI initiatives (e.g. AI users, decision makers, AI strategists, business domain experts), as well as gathering value-related artifacts (e.g. business cases, cost benefit analysis, benefit realization measures and reports) would provide a more holistic view while also unpack the micropolitics of the AI value creation process. In our study, we also did not focus on the unintended consequences of AI applications (Newell and Marabelli 2015)—an aspect that might severely impact value creation while at the same time allows for broader conceptions of value to be taken into account like societal value. Longitudinal case studies might provide richer insights into the unintended consequences of AI applications and their impact on value creation. While our findings are a first step towards understanding AI value creation, future studies could build on the limitations and provide additional contributions in this research area.

3 Paper III

Towards Design Principles for Data-driven Decision Making – An Action Design Research Project in the Maritime Industry

By Tiemo Thiess

and

Oliver Müller

Abstract

Data-driven decision making (DDD) refers to organizational decision-making practices that emphasize the use of data and statistical analysis instead of relying on human judgment only. Various empirical studies provide evidence for the value of DDD, both on individual decision maker level and the organizational level. Yet, the path from data to value is not always an easy one and various organizational and psychological factors mediate and moderate the translation of data-driven insights into better decisions and, subsequently, effective business actions. The current body of academic literature on DDD lacks prescriptive knowledge on how to successfully employ DDD in complex organizational settings. Against this background, this paper reports on an action design research study aimed at designing and implementing IT artifacts for DDD at one of the largest ship engine manufacturers in the world. Our main contribution is a set of design principles highlighting, besides decision quality, the importance of model comprehensibility, domain knowledge, and actionability of results.

Keywords: Data-Driven Decision Making, Design Principles, Action Design Research

3.1 Data-driven Decision Making and its Business Value

Data-driven decision making (DDD) describes organizational decision-making practices that emphasize the use of data and statistical analysis instead of human judgment only (Brynjolfsson et al. 2011). Provost and Fawcett (2013) understand DDD as the outcome of data science, which they define as follows: "Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data" (p.53). Moreover, they state that data science relies on (big) data processing and engineering. So, following Provost and Fawcett (2013), DDD is the outcome of data science, data processing, and data engineering processes.

DDD is rooted in different technical disciplines, such as, decision support systems (Arnott and Pervan 2008; Shim et al. 2002), business intelligence (Chen et al. 2012), data mining and knowledge discovery (Fayyad et al. 1996), and machine learning (Bishop 2006). But to turn data into value, it is equally important to also consider behavioral aspects of human judgment and decision making (Kahneman 2003; Thaler 1980; Tversky and Kahneman 1992). Human judgment can, for example, be affected by cognitive biases. Due to their limited information processing capacities, humans often apply simplifying heuristics for making decisions, especially in situations characterized by high uncertainty (Tversky and Kahneman, 1974). Consequently, human judgments tend to be inferior to formal or algorithmic predictions in terms of predictive accuracy (Grove et al., 2000). Yet, at the same time there is a growing number of critical voices arguing that algorithmic decisions can be subject to biases too (Boyd and Crawford 2012); for example, because they enact simplistic approaches to knowledge creation, are built on an uncritical use of black-boxed assumptions, or lack accountability and transparency (see Latour 1987; Suchman 2002; Winner 1980).

From an economic perspective, there is a growing body of literature suggesting that DDD generates business value. Davenport and Harris (2007), for instance, found a positive correlation between the adoption of analytics in organizations and their annual growth rates (based on a

survey amongst 32 companies). A survey research study by (Brynjolfsson et al. 2011) supported this finding by showing that, amongst 179 surveyed companies, the adoption of DDD leads to an increase in firm productivity of 5-6 percent. Likewise, a recent study by Müller et al. (2018) examining more than 800 firms over a period of seven years showed that the use of big data and analytics is associated with an average increase in firm productivity of about 4 percent, with some industries reaching returns on BDA of more than 7 percent. A similar positive impact of DDD on firm productivity was reported by Wu and Hitt (2016), but they also found that the value of DDD is mainly in enabling continuous process improvements (exploitation) and not in sparking disruptive product or service innovation (exploration). These quantitative studies are further backed up by a large number of qualitative case studies that generally report positive findings about the relationship between DDD implementation and business value (e.g., vom Brocke et al. 2014; Sodenkamp et al. 2015).

To sum up, existing research strongly suggests that DDD generates business value. However, the current body of knowledge on DDD mainly focuses on descriptive and explanatory studies. What is lacking, so far, is prescriptive knowledge on how to design and implement DDD in complex organizational settings. Moreover, there is a lack of research that investigates the role of decision-making processes and human judgment on the outcome of DDD implementations (Sharma et al. 2014).

Against this background, this paper reports the results of an Action Design Research project that was aimed at designing and implementing IT artifacts for DDD at one of the largest ship engine manufacturers in the world. Besides presenting the design of the artifact itself, we formulate a set of nascent design principles for DDD, that can help other researchers and practitioners to implement DDD in comparable settings.

The remainder of this paper is structured as follows: We first provide a theoretical background on the challenges of implementing DDD in organizational settings. We then describe the action design research method in general, before we report on the process and outcome of applying it in our case. In the main part of the paper, we present four proposed design principles and ex-

plain their theoretical and empirical justification. The paper concludes with a short summary and outlook.

3.2 Challenges of Implementing DDD

To discuss common challenges of implementing DDD, we use the data-to-insight-to-decision-to-value conceptualization by (Sharma et al. 2014) as a framework. Even though Sharma et al. (2014) use it to elaborate on a research agenda for creating value from business analytics, we find it particularly suitable as a framework as, in distinction to more established DDD concepts (for instance Shearer 2000), it acknowledges the importance of human judgment and decision-making processes to creating value with DDD. Moreover, it supports our diagnostic that prescriptive knowledge about how to implement DDD to create value is lacking and that without appropriately considering human judgment and decision-making processes already in the design of DDD artifacts, DDD cannot unfold its potential, or even fail in some cases.

Data to Insight

Nowadays, organizations have technologies at hand that enable them to collect, store, manage, analyze, and visualize large volumes of data of varying formats and at increased velocity (Müller et al. 2016; Watson 2014). Nevertheless, as Sharma et al. (2014) point out, "despite the hopes of many, insights do not emerge automatically out of mechanically applying analytical tools to data. Rather, insights emerge out of an active process of engagement between analysts and business managers using the data and analytic tools to uncover new knowledge" (p. 435). One of the most common mistakes in generating new knowledge from data is to start with the wrong initial question, or not having a clear question at all (Leek and Peng 2015). For example, inferential questions are often confused with causal questions, leading to confusion between spurious correlations and real cause-and-effect relationships. Or analysts may confuse exploratory questions with inferential questions, also called "data dredging", or exploratory questions with predictive questions, leading to "overfitting". A way to overcome such pitfalls is to compose multi-disciplinary data science teams that possess not only the required statistical and

computational skills but also the necessary domain knowledge to formulate the right questions and draw valid conclusions from analysis results (Sharma et al. 2014).

Lycett (2013) emphasizes the involvement of human "sense-making" in the process of turning data to insights. Following Lycett (2013), designers of DDD solutions take important decisions regarding what data is selected and what inferences are drawn from the data. Moreover, as designers are human, they are also prone to human biases (Tversky and Kahneman 1974), which affects the insights that are generated and the decisions taken based on them.

Insight to Decision

Research on judgment and decision-making provides strong empirical and theoretical arguments that favor algorithmic or statistical decision making over human judgments, particularly when it comes to complex decisions (Evans 2006; Tversky and Kahneman 1974). For example, a meta-analysis of 136 empirical studies that compared statistical predictions and human judgments in fields ranging from clinical decision-making to economics showed that statistical techniques lead on average to a 10 percent higher accuracy than human judgments (Grove et al. 2000). The superiority of statistical methods over human judgments holds for trained, untrained, experienced, and inexperienced judges (Grove and Meehl 1996). Theoretical explanations for these findings include human biases (e.g., ignoring base rates, failure to take regression toward the mean into account, over-weighting individual factors) and judgment heuristics (e.g., representativeness, availability, and anchoring and adjustment; Tversky and Kahneman 1974).

Yet, despite the overwhelming evidence for the benefits of using insights generated from data to inform decision making, practice shows that new insights are not automatically translated into good decisions. Instead, the conversion of insights into decisions is influenced by a host of psychological and contextual factors (Sharma et al. 2014). For example, due to humans' limited information processing capacities, decision makers tend to satisfice, that is, select a course of action that will satisfy the minimum requirements needed to achieve a particular goal, but which is not necessarily the optimal alternative (Simon 1956).

In addition, organizational decision-making processes and practices can have a strong influence on translating insights into decisions, such as management's inertia in moving towards a data-driven culture or a fragmented use of analytics in single departments instead of enterprise-wide adoption (SAS 2012). Adding to this, survey results by LaValle et al. (2011) and Ransbotham et al. (Ransbotham et al. 2015) suggest that a "lack of understanding of how to use analytics to improve the business" and "turning analytical insights into business actions" are among the top challenges hindering a successful implementation of DDD.

Decision to Value

As mentioned earlier, there exists a growing body of empirical evidence that the implementation of DDD leads to increased organizational performance. However, these benefits are not evenly distributed across all industries and business functions. Müller et al. (2018) showed, for example, that only companies in certain types of industries are able to extract measurable productivity improvements from the use of big data and analytics, and according to Wu and Hitt´s findings (2016), the value created by DDD is mainly exploitative and gained via process optimizations.

One obstacle for turning better decisions into higher value is the observation that it is by no means certain that effective decisions will also be successfully implemented (Sharma et al. 2014). Besides decision "quality" (effectiveness), another important criterion of good decisions is decision "acceptance", that is, the likelihood that stakeholders responsible for the successful implementation of the decision commit to it (Sharma et al. 2014). Prior research suggests, amongst others, that the level of stakeholders' participation in the decision-making process (Hollander 1973) and the comprehensibility of the underlying decision model (Kayande et al. 2009) are factors impacting on decision acceptance – both of which are often not always given in automated DDD processes.

Furthermore, Sharma et al. (2014) argue that even when self-optimizing machine learning algorithms are applied, the outcome of those algorithms still needs to be accepted by human decision makers regarding its validity and usefulness, for instance: "in 'deciding' to deploy them to

run operations in an unguided manner, and in 'accepting' the refinements to the algorithms generated via machine learning as being valid" (p. 436).

3.3 Action Design Research

To develop design principles for how to design and implement DDD in complex organizational settings, we employed Action Design Research (ADR) as a research method. ADR is "a research method for generating prescriptive design knowledge through building and evaluating ensemble IT artifacts in an organizational setting" (Sein et al. 2011, p. 40). The motivation for ADR is to better serve the "dual mission" of Information System Research, that is, to "make theoretical contributions and assist in solving the current and anticipated problems of practitioners" (Benbasat and Zmud 1999; Iivari 2003; Rosemann and Vessey 2008 as referenced in Sein et al. 2011, p. 38). Compared to more traditional design science research methods (e.g., Hevner et al. 2004; Peffers et al. 2007), which are often conducted in the form of stage-gate processes leading to a disconnect between the development of artifacts and their actual application in organizational settings, ADR fully recognizes the role of organizational context in shaping the design process as well as the deployed artifact.

The actual process of ADR consists of four stages, which build on different principles and tasks. The first stage, "Problem Formulation", is based on two principles: "Practice-Inspired Research" and "Theory-Ingrained Artifact". The first principle emphasizes that problems from the field can be knowledge-creation opportunities. Following this, the researcher's intent should not only be to solve a specific instance of an encountered problem, as a software engineer or consultant might do, but to generate general prescriptive knowledge that can be applied to solve the class of problems that the specific problem instance exemplifies. The second principle of the first stage acknowledges that the design and evaluation of artifacts should be informed by existing theory, rather than solely driven by the designer's creativity. In particular, there are three ways of using prior theory in ADR: (1) to structure the problem (2) to identify solution possibilities (3) to guide the actual design. (Sein et al., 2011) This reflects the assumption behind ADR that 'the action design researcher actively inscribes theoretical elements in the ensemble artifact, thus manifesting the theory "in a socially recognizable form"'(Orlikowski and Iacono 2001, p. 121 as

cited in Sein et al., 2011). This, however, constitutes just the first stage of ADR: "[The artifact] is then subjected to organizational practice, providing the basis for cycles of intervention, evaluation, and further reshaping" (Sein et al., 2011, p. 41).

The second stage of ADR, "Building, Intervention, and Evaluation" (BIE), builds upon three principles: "Reciprocal Shaping", "Mutually Influential Roles", and "Authentic and Concurrent Evaluation". (Sein et al. 2011) Reciprocal shaping refers to the complex relations and mutual influences between the designed artifact and its organizational context. The researcher may, for example, use the artifact to gain a better understanding of the organizational environment and then use this increased understanding to refine the selection of design constructs. The principle of mutually influential roles emphasizes the need for mutual learning between the involved roles, being the researcher(s), practitioners, and end-users. These roles, however, can overlap. The principle of authentic and concurrent evaluation points to the key characteristic of ADR that building and evaluation are not conducted in separated stages, but are rather ongoing activities that also involve practitioners and end-users into the design process: "Consequently, authenticity is a more important ingredient for ADR than controlled settings" (Sein et al., 2011, p. 44).

In the third stage, "Reflection and Learning", the researcher moves from building a solution for an instance of a problem to applying that learning to a broader class of problems. The principle "Guided Emergence" describes that the artifact is not just a result of the initial theory-informed design (Stage 1), but of multiple cycles of complex and continuous shaping in the context of the organization (Stage 2), e.g., due to new upcoming requirements or refinements based on insights from user involvement and empirical evaluations. Those refinements to the initial design of the artifact "provide an opportunity for the ADR team to generate and evolve design principles throughout the process" (Sein et al. 2011, p. 44).

The fourth stage, "Formalization of Learning", is based on the principle of "Generalized Outcomes". Because of the described aspect of situated learning, including aspects of organizational change together with the actual implementation of an artifact, the generalization of ADR out-

comes can be tricky. However, to address this issue, it is suggested to generalize the generated knowledge, this is possible on different levels: (1) generalization of the problem of an instance, (2) generalization of the solution instance, and (3) derivation of design principles from the design research outcomes. (Sein et al. 2011)

3.4 Data-Driven Lead Generation in the Maritime Industry

In the following sections, we report from our ADR project of data-driven lead generation in the maritime industry following (Sein et al. 2011) and their suggested ADR steps and principles.

3.4.1 Problem Formulation

3.4.1.1 Practice-Inspired Research

We got the opportunity to work with one of the biggest international engine manufacturers in the maritime industry. The particular department that we worked with is supporting the global aftersales business with data analytics, process, and project capabilities. The department´s technical core is a mature enterprise data warehouse that extracts, transforms, and loads data from multiple sources into a common format and location for analysis by enterprise users. Moreover, the department is responsible for several digitalization projects, amongst those, the implementation of a company-wide CRM system that enables the company to support and optimize sales processes, to store important customer data at one shared location, and finally to become more customer-centric (one face to the customer).

An interesting first diagnostic that informed our conceptualization of a research opportunity is that from the department´s comprehensive portfolio of analytical apps, the apps with the highest usage are those that support and improve an existing business process. In contrast, more explorative apps, which are not embedded in a current or new business process, are those with the lowest usage, even though in the long run they might be much more promising than others. On the one hand, this supports the finding of Wu and Hitt (2016) that the value generated from DDD is mostly exploitative, on the other hand, it shows a need for developing business processes around DDD solutions and, thus, to shift the focus in DDD away from the data-to-insight

process alone to the holistic data-to-insight-to-decision-to-value process (Sharma et al. 2014) in order to increase user adoption and value creation of DDD artifacts.

Furthermore, we found that the company-wide CRM is perceived as a promising and necessary tool to make the company more customer-centric. However, many sales processes are still key-account-driven and not well aligned with the pro-active approach that the new CRM system supports. So, there is a situation in which the system is ready for pro-active sales processes, but the organization needs still time to adapt to this new pro-active approach, especially because the users are partly lacking business processes surrounding the new system and its affordances. Those diagnostics led us to formulate the field problems as follows:

lack of business process embeddedness for low-usage DDD applications

under-utilization of CRM system due to lacking pro-active business processes surrounding it

The resulting initial research opportunity and question was:

How to enable pro-active CRM processes via DDD?

Following the suggestion from Sein et al. (2011, p. 40) to "generate knowledge that can be applied to the class of problems that the specific problem exemplifies", we abstracted the formulated field problems to the class of DDD-value-creation-problems.

3.4.1.2 Theory-Ingrained Artifact

Sein et al. (2011, p. 41) suggest three ways of using theory in the initial design of an artifact: "to structure the problem (…), to identify solution possibilities (…), and to guide the design". In accordance, we choose the conceptualization of (Sharma et al. 2014) as a structural framework for discussing and utilizing theory regarding challenges of implementing DDD into the solution (artifact). Moreover, Shearer´s (2000) cross-industry standard process for data mining was chosen for guiding the design of data science sub-artifacts. Furthermore, Dearden´s (2001) conceptual information-decision-insights-supervision framework (IDA-S) was chosen as a design theory to guide partially automated characteristics of the artifact.

The organizational support for the project was secured by managing expectations and involving stakeholders such as the application manager of the CRM system and a business manager into the design process from the beginning on. Moreover, one of the researchers is working as an industrial PhD at the host company, which helped to anchor the project in the organization.

3.4.2 Building, Intervention, and Evaluation (BIE)

3.4.2.1 Reciprocal Shaping

The general solution understanding was informed by initial design principles of pro-activity, embeddedness, partial automation, and data-drivenness that were derived from the diagnosed field problems and selected theory. The main design objective was to design a DDD artifact that creates a new data-driven and pro-active lead generation process within the CRM system.

In the first iteration, we developed a concept to generate lead events based on predicting upcoming major overhaul events for engines using machine learning algorithms trained on transactional data of spare parts sales. However, we lacked historical data regarding major overhaul events. The reason for this is that in the maritime manufacturing industry, in general, large amounts of data are available, however, on a product or event level, correctly labeled transactional data can be sparse.

In the next instance, we found an alternative approach to generate leads from transactional data. In particular, we found that certain events in the life cycle of ships, such as changes in ownership or upcoming dry dockings of ships, constitute lead events. However, this knowledge is usually not available in digital form but gained through implicit and informal key-account management activities or other forms of direct customer contact, e.g., during a service visit. Yet, we were able to identify an external database of ship registrations, which could be repurposed to extract the required information about lifecycle events by applying several business rules to transform the data. After a successful proof of concept, we worked closely together with practitioners to develop the right business rules and integrate them into the production version of the department's data warehouse.

At this point, we were able to generate initial sales leads based on relevant events in the life-cycle of ships. However, in many cases, the event-driven approach generated simply too many leads to follow up on all of them. As a result, we proposed to prioritize and segment the customer base so that leads can be selected according to metrics of (future) customer behavior, such as their customer-lifetime value (CLV), purchasing patterns, and probabilities to churn in a given future period (see Fader 2012). One of the theoretical ideas behind calculating CLV is "customer centricity", which suggests focusing efforts on the customers with the highest future CLV. The assumption is that it is more rewarding to focus on already strong customer relationships than to try to (re-)launch weak customer relationships (Fader 2012). After exploring different modeling approaches on the transactional customer data at hand in combination with an extensive literature search, we decided on using so-called Bayesian Buy-Till-You-Die probability models for estimating the customer metrics of interest. Amongst the reasons for this choice was that the company operates in a non-contractual market setting, which means that it is not clearly observable when a customer relationship ends and the next transaction occurs (Fader and Hardie 2009), in contrast to, e.g., cellphone subscriptions. Another reason was that hierarchical Bayesian probability models allow for estimating individual-level parameters (Abe 2008; Peter E. Rossi and Greg M. Allenby 2003) and can utilize cohort level information when individual-level data is lacking (Efron and Morris 1977).

After developing a working prototype, we contacted one of the company's regional sales organizations to introduce the initiative and run a pilot of the developed method. The resulting meetings were very insightful especially regarding how to enrich the generated leads with further customer, ship, and engine information, so that they can be represented in the CRM system in a way that the sales responsible can directly take action to follow-up, without having to seek for information elsewhere. In particular, we attached a slide deck to the leads that explains the lead generation campaign, e.g., dry dockings or owner changes, in detail. Moreover, we attached excel reports with further customer and ship insights.

Eventually, we developed a method of a data-driven lead generation that contains five steps. **First**, by applying look-up algorithms to compare the current version of the external ship regis-

tration database with the version from the month before, we create a change log of ship information to identify changes in the ships´ life cycle stages and, in turn, generate initial leads. **Second**, as there can be situations in which there are too many leads in a given month, we calculate CLVs and other behavioral customer metrics for the customers of interest to, for instance, identify the leads for customers with the highest future CLV, or leads for customers that are at risk to churn. **Third**, we enrich the leads with further customer and ship information from the company's data warehouse. **Fourth**, we use a lead uploading template to create and assign the generated leads directly in the CRM system. **Fifth**, we evaluate the performance of the generated data-driven leads via feedback meetings with the sales organizations and via quantitative analysis of CRM data to learn about and improve the quality of the generated leads. The first four steps of the method can be fully automated and implemented into extract transform and load processes (ETL).

3.4.2.2 Mutually Influential Roles

We conducted the BIE cycles following an IT-dominant schema (Sein et al. 2011) in which we were the researchers but also the leading designers and engineers of the artifact. Therefore, we were responsible for the formulation and technical implementation of design principles to ultimately create user-utility via an artifact for data-driven decision support. In this process, we were supported by a design team that consisted of a senior data warehouse engineer and student workers from the aftersales data analytics department. In addition, a wider group of business professionals and test-users from the company was supporting the team with valuable domain knowledge throughout the entire design process.

3.4.2.3 Authentic and Concurrent Evaluation

After the different design instances, the artifact was evaluated with regard to changes to the problem understanding, design principles, the need for further design cycles, and organizational effects. So far, 288 data-driven leads were created in the CRM system from which 73% have been worked on. We also got very positive feedback from the application manager of the CRM system, as one of the key stakeholders:

"It's very interesting to see what scientific theories applied on our data sources can be used for. It has been important for us to include some of the receivers/end-users of the data-driven leads in the process to make it tangible for them and gain from their real-life expertise and not end up with a bunch of leads that only looked promising on paper. Having their stamp of approval is the first step towards a more pro-active sales process and thereby creating additional value. The data-driven leads will be an addition to their work and will save them some time when looking for new leads in the market, these leads come out of the box, being our CRM system."

Also the research question could be addressed with designing and implementing an working DDD artifact that creates pro-active business process by enabling sales responsible to take action without a prior customer inquiry: "We have to search for leads wherever we can, and using the data sources available is a natural next step in a more pro-active sales approach. It's important that we setup an automated process around it and analyze on the outcome of the data-driven leads, to optimize the process over time." (Application Manager CRM System)

3.5 Reflection, Learning, and Formalization of Design Principles

After multiple cycles of building, intervention, and evaluation, we successfully designed and implemented a DDD artifact for data-driven lead generation. This artifact can be used as a tool to develop similar artifacts in many different DDD context. Thus, we abstracted the solution artifact from the maritime industry to the higher class of DDD solutions.

Moreover, we reflected on the changes in problem and solution understanding as well as on design decisions taken and the feedback received from the practitioners. The aim of this phase was to abstract from the specific problems and solutions encountered in the case in order to generate more generic prescriptive knowledge about the design and implementation of DDD. We formulated this knowledge in the form of design principles, following the template proposed by Kruse et al. (2015).

DP 1: Given a lack of proof-of-concept, use theory-based models instead of data-driven machine learning algorithms in order to achieve concrete results.

DP 1 is based on the initial design principle of data-drivenness that was derived from the problem formulation stage. The principle was further shaped throughout the different BIE iterations towards its current formulation. A major design problem that arrived was the choice of a DDD modeling approach that could utilize transactional data for the data-driven lead generation artifact.

Broadly speaking, there are two cultures of using statistical models to gain insights from data (Breiman 2001). The first tries to reconstruct and model the "true" relationships between data inputs and outputs in the form of some mathematical function. Typically, these input-output relations are deductively derived from extant theory, attempt to represent cause-and-effect relationships, and should be interpretable for humans. The second culture treats the process that has generated the data at hand as a complex and unknown black box. Instead of trying to discover the true inner workings of this black box, researchers simply build an algorithm that is able to predict the process' output, given its inputs. The resulting model emerges in a purely inductive fashion, is often based on correlations instead of causation, and is typically incomprehensible to humans.

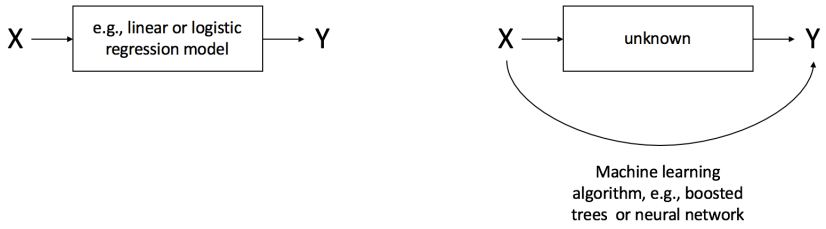

Figure 1. Two cultures of using statistical models

While traditionally the majority of researchers and practitioners followed the theory-based modeling culture (Breiman 2001b; Shmueli 2010; Shmueli and Koppius 2011), the data-driven algorithmic culture became more and more popular with the emergence of big data and the increasing adoption of machine learning in practice. Over the last years, black-box machine

learning algorithms such as gradient boosted trees or neural networks have proved their usefulness when working with large and high-dimensional datasets and outperformed more traditional methods like linear or logistic regression in many of the recent classification and regression competitions. However, according to the no-free-lunch theorem (Wolpert and Macready 1997), there is not one algorithm that fits all problems.

As described earlier, following the trend towards prediction with data-driven machine learning algorithms, we started the project with collecting a dataset comprising engine maintenance events and variables that are potentially correlated with this event. The goal was to create an algorithm that is able to predict maintenance events based on early signals like quotes, orders, or runtimes of engine parts as well as basic customer characteristics (e.g., industry, size). Customers for which the algorithm predicts high probabilities for an upcoming maintenance event are classified as leads and would be assigned to a sales representative for follow up.

However, we soon realized that there were not enough historical observations available in order to train an algorithmic model to the high-dimensional dataset, resulting in overfitting of the model and poor predictive accuracy on test data. Machine learning algorithms have been very useful in the last years for predicting customer behavior in B2C industries characterized by a high volume of transactions (e.g., retail, telecommunications, e-commerce). In contrast, our project is situated in a B2B industry with extremely durable products and a relatively small customer base. Moreover, the transactional data is used as secondary data, only repurposed for analysis. Eventually, the complexity of using machine learning algorithms was too high, as we lacked enough observations of correctly labeled occurrences of major overhauls. After having invested a lot of work and time in this first approach, we learned that it would have been better to have started with a less innovative, but more established and theory-based approach to generating insights from transactional data. This way, stakeholder engagement can be secured by presenting concrete results already at the beginning of a DDD project (quick-wins).

In the following, we decided to utilize an external database of ship registration to create leads based on changes in the life-cycle of ships. To prioritize and segment the leads, we were looking

again for a suitable DDD modeling approach. Based on the learnings from our first DDD modeling iteration and informed by the theorem of Occam´s razor (Blumer et al. 1987), we searched for a DDD approach that constitutes a good trade-off between predictive accuracy and implementation complexity. We then turned to the marketing literature to search for alternative approaches for predicting customers' future purchasing behavior based on customer transaction data. Buy-Till-You-Die models (BTYD; e.g., Schmittlein et al. 1987), an example of theory-based statistical models, and especially those using hierarchical Bayesian models (Abe 2008; Ma and Liu 2007; Platzer and Reutterer 2016), seemed to be particularly suited for our context, because they have been developed for predicting non-contractual customer purchasing (like in our setting), allow individual level parameter estimations, and require surprisingly simple data to be estimated. Only three variables are required for each customer: how many transactions a customer has made in the past (frequency), the date of the transaction (recency), and the monetary value of these transactions. Moreover, due to the possibility of using informative priors, and the utilization of cohort-level information when individual-level data is sparse, hierarchical Bayesian models do not necessarily require big amounts of data to produce good predictive performance (Efron and Morris 1977; Van De Schoot et al. 2015). In addition, BTYD models are based on sound behavioral theory, which enables them to provide useful managerial diagnostics, and have shown excellent empirical performance in the past (Fader et al. 2005).

Eventually, we got the best results with the Pareto/GGG model (Platzer and Reutterer 2016). The dataset was an aggregated version of approximately 500,000 aftersales transactions. To benchmark the model, we predicted the number of future customer transactions one year ahead. Overall, with a mean absolute error (MAE) of 1.2, we got satisfying results. Especially when predicting future transaction for the whole customer cohort, the accuracy was with 93% very good (see Table 1; frequency as target variable had a minimum value of 0.0, a mean value of 1.6 and a maximum value of 93.0 in the validation dataset).

Model	**Actuals / Prediction**	**MAE**
Pareto/GGG	93%	1.2

Pareto/NBD (HB)	78%	1.4

Table 1: Predictive Performance of Pareto/GGG compared to Pareto/NBD (HB)

DP 2: Limit the complexity of models in order to gain support by managers.

Besides predictive accuracy and implementation complexity, comprehensibility is another important feature of any decision support system (DSS), as it increases the trust users put in the outputs of the system and, thereby, drives user acceptance of the system itself (Gregor and Benbasat 1999). Kayande´s et al. (2009) 3-Gaps framework conceptualizes this idea in more detail. It proposes that the sizes of the gaps between the model implemented in the DSS, reality, and managers' mental models influence the performance of the DSS, it's acceptance by managers, and managers' performance.

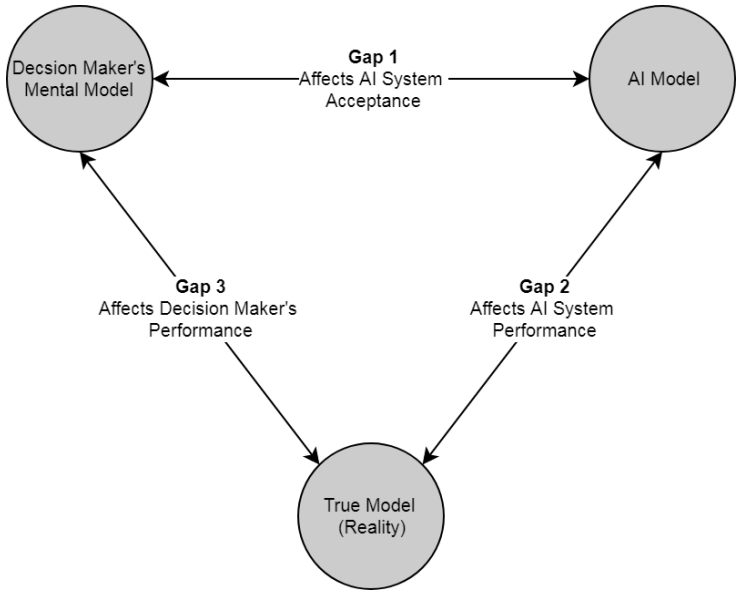

Figure 2. 3-Gaps Framework (adapted from Kayande et al. 2009)

As discussed in DP 1, to provide high predictive accuracy the DSS model must match the true but unknown process that generated the underlying data as good as possible (Gap 2). Likewise, a manager's mental model should be as close as possible to the true model (Gap 3) in order for the manager to make correct decisions (irrespective of using a DSS or not). Our main interest is

in Gap 1. When the manager using a DSS does not understand the logic of the system, the gap between DSS model and the user's mental model is large. Consequently, the predictions provided by the system and the manager's experience and intuition are likely to be in conflict. In such situations, risk-averse decision makers tend to rely on their "gut feeling" instead on the DSS, even if following the advice of the DSS would objectively increase decision quality (Kayande et al. 2009).

The experiences we made over the course of the project support the above outlined theoretical arguments. As described earlier, we started out with using machine learning to classify leads for after-sales service from data about customers' spare parts and consumption profiles. This approach involved black-box algorithms like boosted decision trees and high-dimensional datasets. Although all project members and stakeholders had a strong analytical background, it was difficult for many to comprehend both the inner workings of the algorithms and the meaningfulness of the processed data. Moreover, when working with practitioners from different business domains such as engineering, the mentioned issues became more apparent. For instance, we needed to get detailed technical information regarding labeling major overhaul events from the past, to be able to predict future occurrences of such events. However, as the practitioners were not familiar with machine learning concepts and especially how the many input variables relate to the admired outcome, it was difficult for us to communicate and for them to comprehend what the model was supposed to do.

In contrast, the Buy-Till-You-Die models that we used later in the project were much better received. Although the mathematics behind these models are also complex and were unfamiliar to most project members, they are easier to conceptually understand as they require only three pieces of input information about each customer: their recency (i.e., the time of the last transaction) and frequency (i.e., the number of transactions made in a specified time period), plus the monetary value of the transaction for calculating CLVs. Nevertheless, the limited complexity of the input data creates the need for strong assumptions (e.g., purchase process follows a Poisson process). However, the format of an event log of historical customer purchase transactions and the marketing theory informed assumptions of, e.g., the purchasing and churn process, were

very much in line with the experience and intuition of the involved domain experts, minimizing the gap between the DSS model and managers' mental models.

DP 3: Incorporate domain knowledge into the data-driven decision-making process in order to foster acceptance by managers.

Statistical models, like the Buy-Till-You-Die models we used, induce predictive models of customer purchasing behavior from historical data about past transactions. Apart from this source of information, there, of course, exist human experts who have developed expertise through years of experience in marketing and selling services to customers. In contrast to the statistical models, their domain knowledge tends to be implicit and heuristic in nature, for example, in the form of best practices or rules of thumbs. This knowledge, although it might be difficult to formalize, can still hold valuable information for predicting future customer behavior.

A small but growing stream of research is investigating how human domain knowledge and data-driven predictive models can be combined in order to construct better decision support systems (see, e.g., Dybowski et al. 2003; Sinha and Zhao 2008 for an overview). Sinha and Zhao (2008), for instance, systematically compared the performance of data mining algorithms for credit risk scoring with and without incorporating experts' domain knowledge in the form of rules of thumbs and found that considering domain knowledge significantly improves predictive accuracy. Other researchers improved decision quality by using Bayesian approaches to incorporate prior believes derived from expert judgments into the model estimation process (e.g., Druzdzel and Díez 2003; Langseth and Nielsen 2003).

In our project, we integrated domain knowledge in the form of simple rules capturing experts' experience and intuition into the data-driven decision-making process. More concretely, we interviewed industry experts from the case organization to elicit what types of events at the customer side may lead to a demand for spare parts or maintenance service. We learned, for example, that events such as an upcoming dry-dockings or change in the ownership of a ship increase the likelihood that the owner will order specific spare parts or services. Moreover, we involved regional sales organizations into the development of the method, so that they could

tell us how a lead needs to be represented in the CRM system and what additional customer and ship information is required to take immediate action. This way, eliciting and incorporating experts' knowledge into the artifact, e.g., from the decision makers that the method is targeted towards, also increased their level of participation in and influence on the final design of the DDD process, a key success factor for increasing the acceptance of the final artifact (Hollander 1973).

DP 4: Provide actionable insights instead of quantitative reports in order to increase use by decision makers.

DP 4 is based on the initial theory informed and practice inspired design principles of pro-activity, embeddedness, and partial automation.

Even if a DDD system produces decisions of high accuracy and acceptance, it is not given that end users will follow those decision proposals and take action. In their survey study on "Big Data, Analytics and the Path From Insight to Value," LaValle et al. (2011) highlight that many organizations fail to translate insights into actions because their analytics is too much focused on describing past and current situations and fails to provide actionable and prescriptive information. The authors recommend embedding analytics into operational business processes and users' daily workflows, instead of isolating it in standardized reports that are not accessed on a regular basis. Such a strategy "makes it harder for decision makers to avoid using analytics - which is usually a good thing" (Davenport 2013).

As mentioned earlier, the practitioners of the department made the general observation that analytical applications without a business process surrounding it are used less often, and also that the CRM system was still lacking business processes to further foster usage and value creation. Based on those observations, we decided to push the leads generated by our DDD method directly into sales representatives' daily newsfeed inside the CRM system, instead of building extra reports or dashboards that have to be pulled by them. The processes were designed so that every lead is created as a separate item accompanied with additional information regarding what to do in the form of a clear naming and description text, as well as via an attached slide

presentation regarding the specific campaign. Moreover, based on the meetings with the regional sales organizations, we decided to enrich the leads with further ship and customer transaction information, so that the sales responsibles have all the information that they need for their regular lead follow-up at their fingertips.

The above-described design decisions were based upon the distinction between descriptive ("What has happened in the past"), predictive ("What will happen in the future?"), and prescriptive analytics ("How can we make it happen?"; Watson 2014). By enriching leads identified from ship life cycle events with predictions about future purchasing behavior and descriptive information about campaigns, customers, and ships, we generate prescriptive knowledge that sales responsible can translate into actions.

3.6 Conclusion

Existing research on DDD provides compelling arguments for its value, both on the level of individual decision makers (Grove et al. 2000) and on an organizational level (Brynjolfsson et al. 2011; Müller et al. 2018; Wu and Hitt 2016). Yet, despite recent calls for research, there is a lack of research on how organizational decision-making processes and human judgment shape DDD and on how to implement DDD in complex organizational settings (Sharma et al. 2014). Hence, the goal of this ADR study was to develop practice-inspired, theory-grounded, and field-tested design principles for implementing DDD in the maritime industry, which can help other researchers and practitioners to implement DDD in comparable settings. Besides providing high decision quality, the formulated design principles acknowledge that systems supporting DDD need to be accepted by the involved stakeholders. Hence, our design principles highlight the importance of model comprehensibility, domain knowledge, and actionability of results. Although the proposed principles are inspired by diagnosed problems and grounded in theory and empirical data, due to the situated nature of ADR, we cannot claim that our list of design principles is complete or optimal. Nonetheless, we firmly believe that they represent a valid starting point and can provide the foundations for further research on how to design and implement DDD in complex organizational settings.

Next to the presented design principles, we contribute by abstracting the artifact from a specific data-driven lead generation instance to a tool for generating data-driven leads in many different contexts, thus, we abstract from a specific solution instance to the broader class of DDD solutions. In future research, we attempt to further deepen the analysis of the impact that our designed artifact has on the process from data-to-value. Moreover, we attempt to further shape the designed artifact towards a generalizable tool for creating value with data-driven lead generation.

4 Paper IV

Setting Sail for Data-Driven Decision-Making – An Action Design Research Case from the Maritime Industry

By Tiemo Thiess

and

Oliver Müller

Abstract

To react to new market dynamics, OEM, one of the largest marine equipment manufacturers in the world, was facing the task of transforming its aftersales business from key-account-manager-driven sales processes to more proactive and customer-centric processes. The company had recently implemented an organization-wide customer relationship management (CRM) system to facilitate this transformation. However, the system was not fully used because of a lack of proactive work practices that the system could support. Based on this diagnosis, we developed and applied a method for data-driven lead-generation that uses advanced analytics and automation to leverage internal and external data sources to identify and assess sales leads. To guide the design process, we ingrained the artifact with theory about data-driven decision-making (DDD) and value creation in the form of initial design principles. After several iterations of building the artifact, examining the organizational context, and evaluating the changes that those interventions introduced, we formalized a set of design principles and abstracted

them to the broader class of DDD artifacts, highlighting decision quality but also the importance of model comprehensibility, domain knowledge, and actionability of results.

4.1 Introduction

Marine equipment manufacturers have traditionally focused their efforts on the product development phase in the product lifecycle and the market for newly built main engines and their designs. However, shipbuilders and marine equipment manufacturers have recently suffered from a major drop in demand for newly built vessels and engines because of an over-supply of certain types of vessels in the market. As a result, equipment manufacturers have challenged their traditional business models and shifted their focus from a traditional product-centric approach to a holistic customer-centric approach. In a customer-centric approach, the aftersales phase in the product lifecycle offers considerable potential for innovation. In the approximately twenty years of marine engines' lives, manufacturers generate most of their earnings from sales of spare parts and services like maintenance, repair, and overhaul. As a result, the market for aftersales products and services is much more competitive than the market for newly built engines, in part because the barriers to entry are much lower, as marine engines usually do not require original spare parts or service from only the engine producers.

Against this background, we started an action design research (ADR; Sein et al. 2011) project at OEM (an alias is used for pseudonymization), one of the biggest marine equipment manufacturers in the world. OEM had recently implemented a company-wide customer relationship management (CRM) system to facilitate the transformation from a product-centric to a customer-centric approach in their aftersales business. The CRM system is a promising and necessary tool for this endeavor, but OEM's sales processes had been predominantly key-account-driven and were based on pull mechanisms. Therefore, the existing sales processes were not well aligned with the CRM system's functionalities, which are intended to afford proactive (rather than reactive) sales practices based on a concept of lead and opportunity management. This assessment led us to formulate our first problem diagnosis:

- Under-use of CRM system because of a lack of proactive business processes

We recognized this problem diagnosis as a knowledge-creation opportunity for generating design theory about information systems, particularly about data-driven decision-making (DDD) artifacts. DDD describes organizational decision-making practices that emphasize the use of data and statistical analysis instead of relying only on human judgment (Brynjolfsson et al. 2011). Existing research strongly suggests that DDD improves the quality of individual decision-making and generates business value at the organization level (Brynjolfsson et al. 2011; Müller et al. 2014; Müller et al. 2018). However, the current body of knowledge on DDD lacks prescriptive knowledge on how to design and implement it in complex organizational settings. According to Sharma et al. (2014), DDD artifacts do not create value simply by being applied; their output must be further processed into actionable judgments. The conversion from insights to decisions to actions and business value appears to be especially challenging (Sharma et al., 2014) in part because the implementation of DDD, unlike large enterprise systems, is often not accompanied by change-management activities (Hollander et al. 1973; Kayande et al. 2009; SAS 2012; Ransbotham et al. 2015), perhaps because of an over-emphasis on the extraction process of insights from data in scientific and industry publications. Therefore, there is a need for research that investigates the role of decision-making processes, human judgment, and change management in generating the outcome of DDD implementations (Sharma et al. 2014).

Another diagnosis that resulted from the initial problem evaluation was that, among the aftersales organization's portfolio of analytical apps, the apps with the most use were those that support and improve an existing business process. In contrast, more explorative apps, which are not embedded in a current or new business process, had the least use, even though, in the long run, they might be much more promising than others. This observation supports Wu et al.'s (2017) finding that the value generated from DDD is mostly exploitative. There is a need to develop business processes around DDD solutions, thus shifting the focus in DDD away from the data-to-insight process alone to the holistic data-to-insight-to-decision-to-value process (Sharma et al. 2014) to increase user adoption and value creation of DDD artifacts. This observation led us to formulate our second problem diagnosis:

No business process embeddedness in low-use DDD applications

Based on those diagnostics, we formulated the following research question:

How can proactive CRM processes be enabled via DDD?

Against this background, this chapter reports the results of a multi-year ADR project in which IT artifacts for DDD were designed and implemented at one of the world's largest ship engine manufacturers. [3]

4.2 The Context

OEM is the power engineering arm of Engineering AG (an alias is used for pseudonymization). OEM is primarily a manufacturer of large bore 2- and 4-stroke diesel engines that can be used in marine vessels and power plants, but it also produces gas engines, dual fuel engines, and turbomachinery. OEM also provides power-generating 4-stroke engines, propulsion solutions, and turbochargers for marine vessels.

OEM has a global aftersales organization that offers original spare parts and ad hoc and contract-based operation and maintenance service via a worldwide network of local sales companies. The ADR case presented here is situated in the marine engine aftersales-service part of OEM. The marine engine aftersales market is a complex and dynamic market that depends heavily on the number of newly built high-sea ships. Overcapacity in the supply of ships currently pressures the market, causing OEM's aftersales-service business to grow in importance, as ship owners prefer exploiting and upgrading their existing fleets to ordering new ships, so

[3] Earlier stages and iterations of the artifact of this ADR study have been reported in an unpublished master's thesis and, with a focus on developed design theory, in the proceedings of the European Conference on Information Systems (Thiess and Müller, 2018).

customers become more interested in long-term service agreements, which are more like subscription-based business models.

The growing importance of the aftersales service business forced the company to undergo significant changes in their sales processes, which have been built on long and close customer relationships with a strong focus on key-account-management practices. To leverage the potential of the aftersales market, OEM wants to enhance its traditional sales approach, for instance, via a digitization initiative. A cornerstone of this initiative is to create more proactive sales processes and services that are built on in-depth knowledge of their customers' needs.

The market for newly built ships is highly transparent. Most major yards are known, as are the main competitors in this market. Moreover, new ships have to be registered with the International Maritime Organization (IMO), which requires detailed information on the type of ship, the ownership structure, and the type of engine. In this market, OEM's large licensee partners are valuable business partners that produce most of OEM's engines, for which OEM grants them access to its state-of-the-art and continually improved designs. In contrast, in the aftersales business, it is often unclear who the competitors are, as marine engines can be serviced by many, often small, companies that do not necessarily have to use original spare parts for repairs. In addition, the licensee partners that are producing engines for OEM are themselves among OEM's strongest competitors in the aftersales market, as they can use OEM's regional network and customer relationships. What's more, as most engines are sold without long-term service agreements, the aftersales market is characterized by high uncertainty, especially with regards to CRM and customer life-cycle management. For instance, it is difficult to define when an aftersales customer relationship starts, when it ends, and what sales volume it will generate.

Current sales practices in the aftersales market are often based on recommendations regarding how many spare parts an engine should use for a certain number of running hours to guarantee high performance. However, information about an engine's running hours are not always easy to obtain, as the engines are owned by the ship owners, and information on running hours is retrievable only from particular customers during periodic on-board service visits. Another

problem is that service visits are documented in text form, so from a data perspective, they are not in an easily analyzable format. Consequently, the information from these reports is not stored in the organization's data warehouse, creating situations in which sales and engineering professionals lack a coherent overview of customers because they have to deal with several unconnected lists and reports in varying formats. Because of the challenges in obtaining an engine's running hours, the ship's age, together with an expert estimation, is typically used as a proxy for running hours. This approach is intuitive and comfortable but not always accurate, as it does not incorporate information about, for example, downtime, breakdowns, or dry-dockings. Therefore, in the context of OEM's digitization initiative, projects were initiated to improve the quality of data for engines' running hours, thereby facilitating more accurate product-lifecycle management. One of these projects uses satellite data on ships' positions to estimate running hours, while another builds an Internet-of-Things (IoT) infrastructure that facilitates the collection and transmission of running hours and other performance data via sensor networks. The goal is to monitor running hours and other performance indicators of connected equipment centrally so OEM's technical experts can optimize engine operations and maintenance.

For the large 2-stroke main engines, another driver of the aftersales business is the dry-dockings of ships that occur approximately every five years because they are required by international shipping societies before certifications can be granted or renewed (International Maritime Organization 2015). Dry-docking may also be done for cleaning, hull maintenance, damage repair, and other unplanned events. As most systems and engines on board are turned off during dry-docking, it is a perfect occasion to perform minor and major overhauls on the engines. However, as with the running hours, when such a dry-docking is taking place is not always easy to know. External databases contain data about the approximate date of the next dry-docking that the registration societies require, but where the dry-docking is taking place and the date on which it actually takes place is difficult to determine.

These problems are just two examples of challenges that companies in non-contractual market settings face. These settings are usually characterized by a high degree of uncertainty regarding

customer behavior and life-cycle management (Fader and Hardie 2009), and when a customer makes its next purchase is usually unclear. Therefore, OEM is focusing on improving its long-term-service-agreement business in a gradual transformation from a mostly non-contractual to a mostly contractual setting. These developments go along with transforming the overall aftersales business model from a product-focused to a service- and customer-oriented model. As a result, increasingly fine-grained and high-quality customer data will become available to facilitate the delivery of smarter services (Beverungen et al. 2017), which is an opportunity to improve products and CRM.

CRM systems and proactive sales processes are today broadly applied in business-to-consumer (B2C) industries like private banking, but they are not used extensively by companies that operate in complex business-to-business (B2B) industries. Among the first implications of OEM's new customer-centric digitization initiative is the recent introduction of a company-wide CRM system, a unified platform that will help to align sales processes with improved product and customer lifecycle management, more customer centricity, and proactivity. The system gives OEM the opportunity to improve how it addresses its mostly non-contractual customer base and supports improvement in the company's understanding and identification of customers with a high potential need for long-term service contracts. Thus, it supports turning non-contractual customers into contractual customers. Relationships with contractual customers also benefit when OEM's abilities to prevent churn and generate up- and cross-selling effects improve. However, the platform is still in its implementation phase, so its full potential has not yet unfolded. Many existing business processes still need to be adapted to the new sales tool, and in many cases, entirely new proactive processes are required if the platform's capabilities are to be used fully.

Finally, it is necessary to define OEM´s aftersales customers clearly. OEM's customer can be a shipowner or, more often, a technical manager of a ship who is authorized to order spare parts and other aftersales services (also called a motor manager). However, when OEM implements more advanced CRM practices, it can be helpful to define the customer as a particular ship or even as an engine on a ship, especially when sales activities need to be closely aligned with the

product's lifecycle. The purchase of a certain combination of spare parts could indicate, for instance, a particular event in an engine's lifecycle. Such insights are much harder to gain when the broader definition of a customer as the technical manager of a ship is used because a technical manager is often responsible for not just one but a whole fleet of ships.

4.3 The Journey

4.3.1 The Action Design Research Process

To generate prescriptive design knowledge in the form of design principles, we employed ADR as a research method. ADR is "a research method for generating prescriptive design knowledge through building and evaluating ensemble IT artifacts in an organizational setting" (Sein et al., 2011, p. 40). ADR combines aspects of action research and design science research (Purao et al. 2010). In particular, the action-research-related concepts of diagnosing field problems, planning action, taking action, and evaluating the effects of the actions taken to specify general learnings (Susman and Evered 1978) are reflected in the ADR method that Sein et al. (2011) propose. Moreover, and in contrast to traditional action research, ADR emphasizes the intervention into an organizational context via designing IT artifacts through an iterative building, intervention, and evaluation stage that adopt, for instance, the concepts of design cycles and rigor cycles from design science research (Hevner 2007).

The ADR method has four main stages (Figure 1). The first stage, the problem-formulation stage, is based on the principles of practice-inspired research and theory-ingrained artifacts and encourages researchers to identify or diagnose field problems and define them as knowledge-creation opportunities and, in particular, as opportunities to develop design theory in, for instance, the form of design principles. Sein et al. (2011) propose three ways of using prior theory in ADR: to structure the problem, to identify solutions, and to guide the actual design using design theories. The concept of ingraining IT artifacts with theory in ADR stems from the idea that IT artifacts are socio-technical assemblages and that researchers can manifest theory by embedding it in an artifact so it can be recognized in a social form (Orlikowski and Iacono 2001).

The second stage, the building, intervention, and evaluation stage, with its principle of reciprocal shaping, acknowledges that IT artifacts shape and are shaped by the context in which they are applied. Another principle is mutual influence and learning among the roles involved in the design process, including researchers, practitioners, and end-users. The third principle of the building, intervention, and evaluation stage is authentic and concurrent evaluation of the artifact throughout the design process (Sein et al. 2011). In this regard, the design evaluation process proposed by ADR may differ from design science research, as in the latter, the relevance, design, and rigor cycles are more separated and often follow a traditional stage-gate approach (Hevner 2007). As a result, ADR focuses on keeping the artifact construction process as authentic and coherent with the design context as possible: "Consequently, authenticity is a more important ingredient for ADR than controlled settings" (Sein et al. 2011, p. 44). The third stage, the reflection and learning stage, is based on the principle of guided emergence and consists of reflecting on refinements of the problem and solution that are visible in the shape and state of the artifact. This stage enables the researchers to adapt and change initial design principles and to recognize newly emerging design principles (Sein et al. 2011).

Finally, in the fourth stage, what has been learned is formalized as generalizable outcomes. This can be difficult due to ADR´s situated nature. Sein et al. (2011) suggest three ways to generalize outcomes in ADR: the generalization of the problem instance, generalization of the solution instance, and derivation of design principles from the design research outcomes (Sein et al. 2011). In our case, we related the emerging design principles to existing theory, thus generalizing our ADR outcomes by means of abduction.

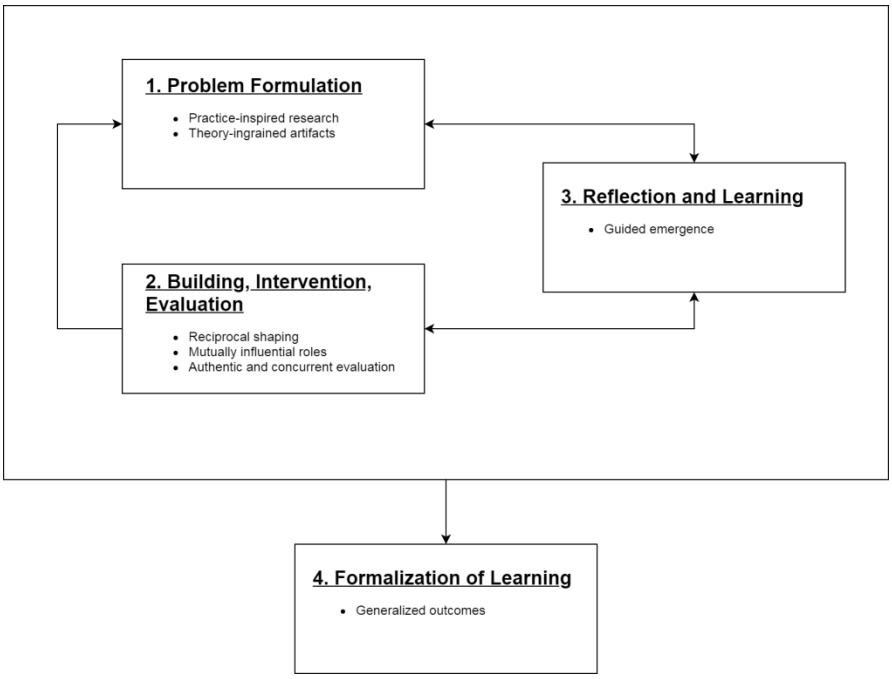

Figure 1. The Action Design Research (ADR) method (Sein et al., 2011)

4.3.2 Our Journey

The ADR team consisted of the authors and a group of practitioners from OEM's aftersales analytics department. One of us, who had been working in the department for around eighteen months when the ADR project started, was employed as an industrial Ph.D. fellow at OEM. The practitioners included the application manager of the CRM system, the department manager, a senior data warehouse architect, and other data and business analysts from the department, who we occasionally involved. Besides the core ADR team, the end-users were valuable contributors of knowledge during the design process. Following the ADR methodology, we, as the researchers, were involved during all iterations and stages of the ADR process, from defining the problem to building and evaluating the artifact to developing and formalizing the design principles. The practitioners were involved primarily in building and evaluating the "alpha versions" of the artifact but also in supporting design decisions with their domain knowledge (Figure 2).

We started our journey by analyzing the current situation at OEM to gain an understanding of the field problem. During the project, we were situated in OEM's aftersales business intelligence

processes and analytics department, which is built around a mature data warehouse that builds the basis for a broad portfolio of analytical applications. The department is responsible for extracting, loading, and transforming transactional data from the company's enterprise resource planning (ERP) systems into ready-to-analyze multi-dimensional data models. Moreover, it is developing and frequently updating analytical reports for a broader audience of business users.

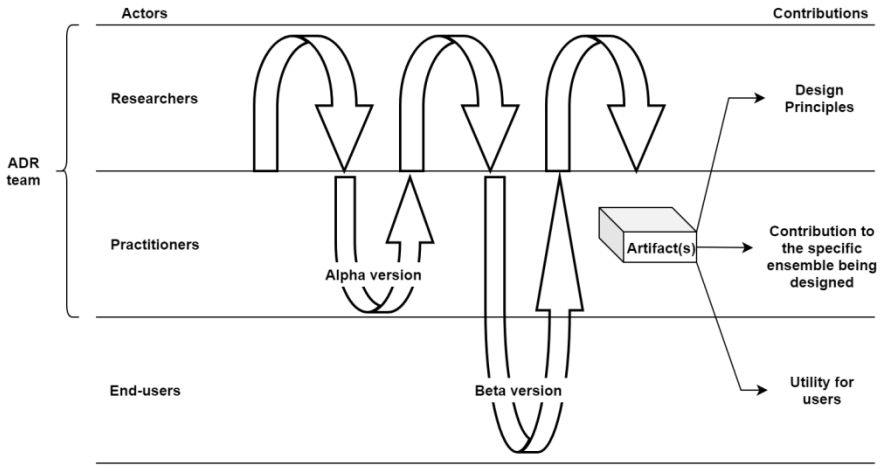

Figure 2. ADR Team, journey, and contributions (Sein et al., 2011)

A diagnosis that guided our initial understanding of the problem was that, in the department's portfolio of analytical applications, applications that directly enhanced a business process or led to creating and implementing an entirely new business process were the most frequently used and most successful ones. In contrast, more innovative applications, which often showed great potential but neither supported existing processes nor straightforwardly showed how a new process could be created around them, were less frequently used and less often successful. This observation is in line with Wu and Hitt´s (2016) findings that the value created from analytics is primarily exploitative and only to a lesser degree exploratory. Therefore, we concluded that in order to generate sustainable business value from DDD applications, new business processes have to be developed alongside the actual DDD applications.

Moreover, we found that the new CRM platform had considerable potential for transforming sales processes to proactivity and customer-centricity. However, many of the existing sales pro-

cesses were key-account-manager-driven, so they did not align well with the platform's capabilities. This observation led us to our second problem diagnosis: Even though the platform was ready to use from a technical perspective, it still lacked users, content, and proactive sales processes.

Based on our understanding of the problem, we evaluated several theories that could support our design process. We selected the cross-industry standard process for data mining (Shearer et al. 2000) to guide the design of the DDD artifacts and the information-decision-insights-supervision framework (IDA-S; Dearden 2001) because of the principle of partial automation that we intended to incorporate into the artifact. We also chose Sharma et al.'s (2014) data-to-insight-insight-to-decision-decision-to-value conceptualization as a structural framework for integrating theory regarding the challenges of implementing DDD in the solution.

Our general understanding of the solution was informed by the initial design principles of proactivity, embeddedness, partial automation, and being data-driven that were derived from the diagnosed field problems and the selected theory. The main design objective was to design a DDD artifact that creates a new data-driven and proactive lead generation-process within the CRM system.

In the first iteration, the alpha version, we intended to generate data-driven leads through detecting and predicting significant events in the life cycle of an engine by relating a customer's spare parts transaction history for a particular engine to the recommended amount of spare parts consumption according to OEM's engine manuals. We received positive feedback when we presented this approach to senior managers at OEM. However, the project was complicated and required expertise from many stakeholders. Because of the complexity of the predictive models that were based on black-boxed machine-learning algorithms, it was difficult to explain the model's inner workings. As a result, and because of the project's overall complexity, we chose to look for a more generic and versatile approach for generating data-driven leads while keeping the already developed alpha version (Figure 3) of the artifact in our back pockets for future iterations of the project.

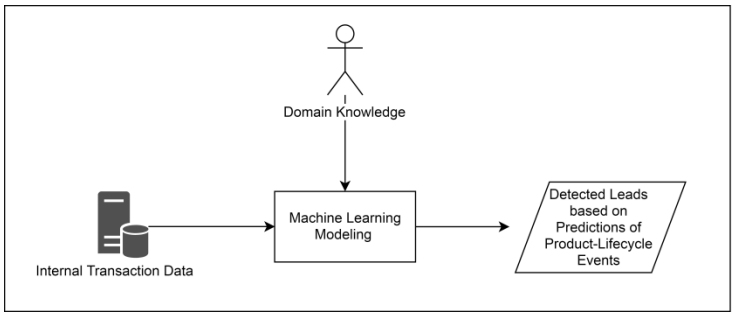

Figure 3. Data-driven lead-generation artifact after the first iteration

In the second iteration, we revised the initial artifact and developed an operational pipeline for data-driven lead generation that is not limited to a specific type of lead (Thiess and Müller 2018). Specifically, we wanted to use a wide variety of data sources that can be connected to the existing data warehouse so we can train and automatically deploy data models on it to generate sales leads. However, to avoid too many leads being created automatically, we looked into the marketing literature to find ways to segment and prioritize customers and leads. We chose to train so-called buy-till-you-die models (BTYD) on customer data to calculate customers' future lifetime values (CLV). Such models enable segmenting and scoring of generated leads based on the predicted CLVs of the concerned customer base, so that, for instance, one can prioritize the top ten leads of customers with the highest future CLV. At this point, the focus of the artifact was on the lead-generation process and did not specify how the leads would be transferred to the CRM system and assigned to a responsible salesperson (Figure 4).

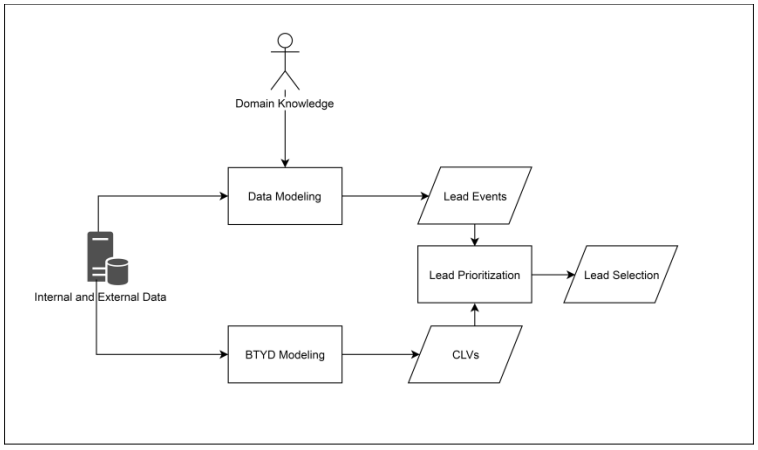

Figure 4. Data-driven lead-generation artifact after the second iteration

In the development of the third iteration, the beta version (Figure 5), we reached out to potential end-users of our artifact. In particular, we scheduled meetings with OEM's local sales companies to get feedback on the current version of the data-driven lead-generation artifact. In particular, we gathered information on the types of lead events that could be extracted from data. As a first lead campaign, changes in the ownership of vessels were identified as lead events that have a high potential to be converted into aftersales business for OEM. In addition, OEM's main aftersales customers are the technical managers of ships. The engines on those ships determine OEM's aftersales market; thus, when a ship's technical manager changes, the customer's relationship with OEM may change too. After a change in ownership, OEM might want to contact the new technical manager of a ship to ensure that the existing service relationship with the affected ship will continue. Changes in technical management are also good opportunities to re-evaluate the customer relationship and seek cross- and up-selling business.

To detect technical management changes automatically, we used an external database of worldwide high-sea ship registrations that is maintained by the international ship registration societies. To that point, OEM's sales professionals could get information regarding changes in the management of ships only via occasional talks with customers or updates of OEM's master data in the standard ERP systems. However, it can take months for this master data to be updated, and even if it is updated, a sales professional must still actively search for the information. To make this process more proactive and faster, we proposed keeping a change log of the technical management registrations in the database. This way, every change in the technical management of ships can be detected automatically via business rules and similarity-matching algorithms. Meetings with sales professionals uncovered an additional need for customer data to enable immediate decision-making and action-taking. In particular, the sales professionals did not want to open multiple IT systems to look for information that they require in order to follow up on leads. In reaction to this, we created two additional reports and attached them to the leads. The first report showed the quote and order history of a particular ship, and the second report contained metadata about the ship and its installed engines.

Another important design question was how to present the leads so they are usable in the CRM system. Because of its potential for automation, we chose an approach based on XML templates that are filled with data about the generated leads. The leads are then automatically assigned to a sales team and uploaded in bulk directly into the system.

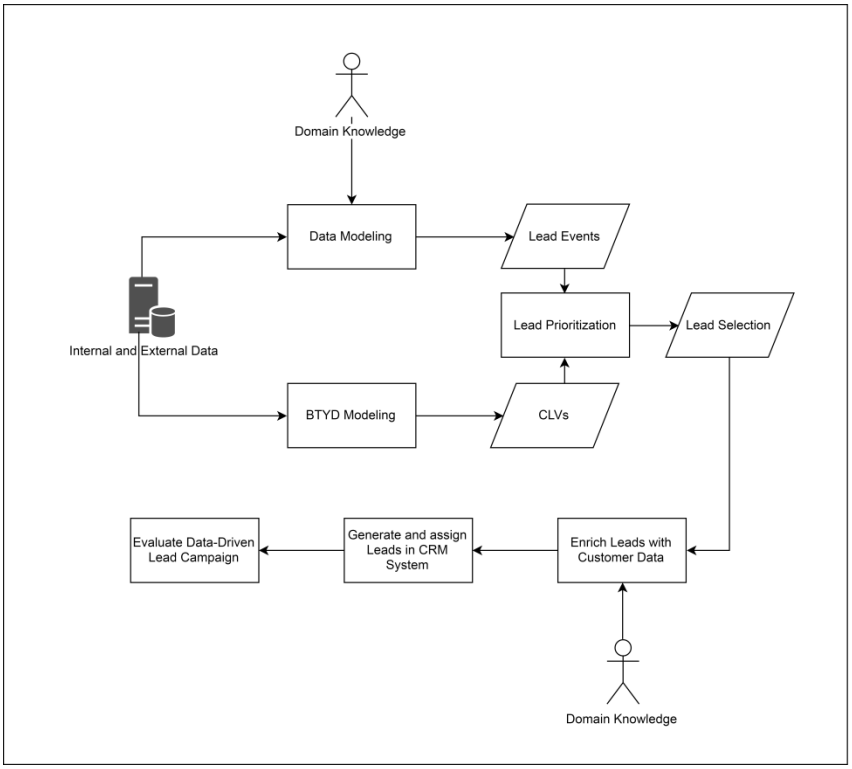

Figure 5. Data-driven lead-generation artifact after the third iteration

In the fourth iteration, we abstracted the instantiation of a data-driven lead-generation pipeline to a method for generating data-driven leads in many contexts. This design iteration was informed by the experience that we gained from the design journey and from the conceptualization of a data-driven decision-making process by Sharma et al. (2014). The fourth iteration, the final solution artifact, is described in detail in the next section.

4.4 The Results – Data-Driven Lead Generation

4.4.1 Presentation of the Artifact: The Data-driven lead-generation artifact

This section introduces the final artifact and explains all of its parts in detail. The key result of the design journey is manifested in the data-driven lead-generation artifact depicted in Figure 6. It is a method to generate data-driven lead pipelines that can flexibly accommodate various kinds of lead events and business contexts. The final artifact contains eight steps for creating business value via operational data-driven lead pipelines. The artifact is constructed as an iterative method that allows the user to fall back to prior steps when necessary. Steps 1 through 7 are the core of a data-driven lead-generation pipeline and account for a full data-driven decision-making process. Step 8 primarily evaluates whether a data-driven lead-generation pipeline should move from a beta state to a fully operational one. However, even after a pipeline is fully operational, it is advisable to apply Step 8 periodically to re-consolidate and evaluate the success of the pipeline.

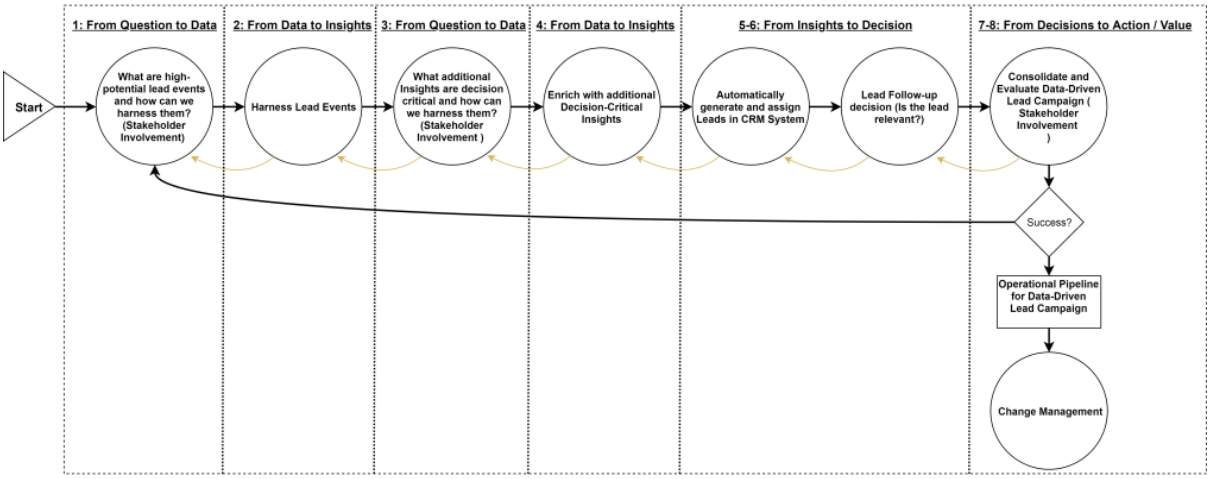

Figure 6: The final data-driven lead generation artifact

The design of the artifact was informed by a generic DDD process that we built based on the terminology Sharma et al. (2014) introduced. The process consists of four main elements:

1. **From Question to Data:** This element defines an initial analysis question, which also entails planning and selecting an algorithmic analytics approach. Then data on which the planned analytics (algorithms) can be applied are made accessible (see Shearer et al. 2000, and Leek and Peng 2015, for a broader description of this element).
2. **From Data to Insight:** This element refers to the application and technical execution of the planned analytics instance (Sharma et al. 2014; Thiess and Müller 2018).

3. **From Insight to Decision:** This element refers to the human-cognition-based decision-making process (Sharma et al. 2014; Thiess and Müller 2018).
4. **From Decision to Action / Value:** This element establishes measures to implement decisions from which to create continuously positive effects (Sharma et al. 2014).

Overall, the final artifact has eight steps that are explained in the following sections.

Step 1: What are high-potential lead events? (Stakeholder involvement)

In Step 1 of the data-driven lead-generation artifact, stakeholders are involved in determining high-potential lead events for a given business context. An important stakeholder group to involve in this step is the potential users of a data-driven lead pipeline—likely those responsible for sales. When the capabilities regarding data-driven lead generation are already established, involvement can also be triggered by stakeholders, but sufficient change management (Kotter, 1995) is necessary to reach that point. In Step 1, high-potential lead events are identified by defining an initial analysis question, planning and selecting an initial analytics approach, and thinking about data sources that may be accessible to the analytics team. From an organization's perspective, data sources are internal and external. Internal data are usually accessible via ERP systems and data warehouses but can also be, for example, SharePoint lists and text documents. External data can be accessed from public application programming interfaces (API). For instance, general economic and financial data can often be retrieved via APIs. Depending on the maturity of an organization's analytical capabilities, making the data for lead events accessible can be challenging, especially if the data quality is low and there is no master data management in place (Wagner and Hogan 1996).

Step 2: Harness lead events

After detecting high-potential candidate leads, it is necessary to determine whether they can be harnessed algorithmically in a data-driven way such that the lead events are technically accessible. This step is not always straightforward; if a high-potential lead cannot be retrieved, users of the method can go back to Step 1.

External and internal data sources must usually be algorithmically processed and analyzed if they are to generate useful insights. The kinds of algorithms required can vary and depend on

the particular lead event as well as on the organization's analytical capabilities. The main types of algorithms and insights generated by those algorithms are (Watson 2014):

- **Descriptive** (e.g., summary statistics, grouping, aggregation)
- **Explorative** (e.g., clustering, dimensionality reduction, visualizations)
- **Predictive** (e.g., regression, classification, time-series analysis)
- **Prescriptive** (e.g., optimization, simulation)

Steps 1 and 2 together cover the elements from question to data and from data to insights in the DDD process and reflect a full data-mining and analytics sub-process. We suggest using the cross-industry standard process for data mining (CRISP-DM; Shearer et al. 2000) for guidance in undertaking Steps 1 and 2.

Step 3: What additional insights are decision-critical? (Stakeholder involvement)

After high-potential lead events have been identified and it is determined that they can be harnessed algorithmically in a data-driven way, relevant stakeholders should be involved again to determine what additional decision-critical insights are needed to take immediate action on a generated lead. For instance, when the lead event alone is not sufficiently prescriptive, we suggest enriching the lead pipeline with further decision-critical insights to support sales representatives in their decision-making and action-taking to create value (e.g., closing a deal).

The level of uncertainty and responsibility for decisions and actions that a sales representative has depends highly on the kind of insights on which the leads are created and with which they are enriched. In the case of descriptive insights, sales professionals have a comparatively high responsibility for the eventual decision and the following action, as there is considerable uncertainty involved about why a customer shows particular characteristics, what the customer may do in the future, and what appropriate (logical) action to take.

The overall objective of Step 3 is to think of and plan for the additional insights that could support sales employees in their decision-making and action-taking. Thus, Step 3 is similar to Step 1, but while Step 1 is concerned with determining the initial lead event and is the basis for the whole lead pipeline, Step 3 is concerned with determining additional decision-critical insights to remove uncertainty from the decision-making process.

Step 4: Enrich leads with additional decision-critical insights

In Step 4, additional decision-critical insights are technically integrated into the data-driven lead pipeline. Step 4 is similar to Step 2 but is focused on additional decision-critical insights instead of the primary pipeline of lead events. Like Steps 1 and 2, Steps 3 and 4 cover the elements **From Question to Data** and **From Data to Insights** in the DDD process, so they require deploying complete data mining and analytics sub-processes. However, unlike Steps 1 and 2, Steps 3 and 4 often entail using a couple of data mining and analytics sub-processes. For instance, reducing uncertainty about an effective lead follow-up may require a combination of descriptive, explorative, predictive, and prescriptive insights.

Step 5: Automatically generate and assign leads in the CRM system

In Step 5, leads are generated in the CRM system and assigned to the right person by notifying them directly when new data-driven leads are available. Step 5 may require connecting and integrating operational sub-processes to join the high-potential lead events with additional decision-critical insights. Moreover, procedures should be put in place to automate the generation and assignment of leads as much as possible, and as long as it makes sense. Here, a trade-off between automation and customizability should be made, especially in the beta phase of a data-driven lead pipeline, as it can be advisable to keep the setting flexible and allow for quick adjustments based on stakeholder feedback.

Step 6: Lead follow-up decision

Step 6 contains the core element of the decision-making process: **From Insights to Decision**. First, sales employees decide, based on the insights provided, whether they will accept an assigned lead or not. A reason for declining a lead could be that the employee already knows about the lead event and has already acted or that the lead is incorrectly assigned, in which case, the employee should delegate the lead to the correct person or inform the project team. However, once sufficient data quality is ensured, it is expected that most generated leads will be accepted.

After a lead is accepted, the sales employee must decide how to act based on the insights provided. The effects of the decision can be either direct or indirect. For instance, the information could be used to contact the customer right away, which is a direct effect of the decision, or the information could be used to change the general sales strategy for the customer, which is an indirect decision effect. Overall, the greater the certainty in a decision, the more likely it is to have a direct effect. In data-driven lead generation, the level of uncertainty is influenced by the kind of insights that are provided to the decision-maker. For example, the insight that a technical manager has changed is descriptive, while additional decision-critical insights regarding the expected future transactions and churn probability of a customer are predictive. Following this, the lowest level of uncertainty is likely reached when prescriptive insights are provided.

Step 7: Change management

Step 7 addresses change management. In particular, we suggest following Kotter's (1995) eight steps to transforming an organization. The change endeavor here focuses not just on the implementation of a data-driven lead-generation pipeline but also on helping stakeholders to understand the basic principles of DDD first. Moreover, the organization's own mindset must often be changed to accept using data-driven and proactive approaches. Kotter's (1995) eight steps appear to be particularly suitable for this purpose:

1. Establish a sense of urgency.
2. Form a powerful guiding coalition.
3. Create a vision.
4. Communicate the vision.
5. Empower others to act on the vision.
6. Plan for and create short-term wins.
7. Consolidate improvements and produce still more change.
8. Institutionalize new approaches.

Step 8: Consolidate and evaluate the data-driven lead pipeline (Stakeholder involvement)

In Step 8, the data-driven lead pipeline is consolidated and evaluated after the beta version has been in use for some time. Stakeholders should be involved again to determine whether the pipeline has been successful, and based on this evaluation, measures can be taken to improve the pipeline or, if it is deemed unsuccessful, to discontinue it in the beta state. However, even an

unsuccessful pipeline is valuable for future pipelines, as parts of it, such as the additional decision-critical insights, can apply to many business problems and contexts in an organization.

4.4.2 Application of the Artifact: The technical manager change pipeline

The following sections describe a concrete application of our artifact, the data-driven lead generation artifact, using the example of the technical manager change pipeline.

In Step 1, we received the information that changes in the technical management of a ship constitute high-potential lead events. We interviewed experienced sales managers with a sound understanding of the dynamics of the marine engine aftersales business. We then contacted potential users (i.e., sales employees) to determine whether changes in a ship's technical management are appropriate lead events in their consideration. As a result of this first stakeholder involvement, it became clear that the sales employees did not have a straightforward process for retrieving information regarding changes in ships' technical management, and that there were no well-defined proactive processes for collecting and working with such information. Instead, sales representatives sometimes received information regarding changes in technical management from customers directly during ordinary sales interactions. Overall, the sales employees stated that the insights regarding recent changes in technical management would be valuable for several reasons:

1. as a conversation starter to contact the customer proactively
2. to learn more about the current customer base and the ships in the territory for which they are responsible (defined by the country in which a technical manager is registered)
3. to update a customer's metadata

These talks led us to conclude that changes in technical managers are high-potential lead events. The next task was to evaluate how and based on what data sources insights regarding such changes could be harnessed. The plan was to investigate the possibility of using external sources to retrieve the data by simply querying it or by using a look-up algorithm to gain the desired insights.

In Step 2, we created the connection to an external data source that contains metadata of ships that are frequently updated with data from the international shipping registries. At first, it looked like data regarding changes in technical management would be comparatively easy to access by querying a column of the dataset that indicates on what date the value of the technical management field for a specific ship was changed. However, during the validation of this initial assumption with the data provider, it became apparent that the column could not be used. Therefore, we chose another approach that consisted of, firstly, creating a log of old versions of the dataset, secondly, creating a script to compare the current value of the technical management field with its latest predecessor, and finally, to trigger an event if the value changed.

In Step 3, we scheduled meetings with sales representatives and the rest of the ADR team to determine what additional decision-critical insights were required. The results of this second stakeholder involvement were to add an additional sales report and a ship and engine report to the leads in the CRM system. This was done to help sales employees to gain a quick overview of the existing and past relationships and interactions with a particular ship and customer. After involving a business manager in the design process, other decision-critical insights in the form of dates of upcoming dry-dockings were identified and added to the reports, along with key customer metrics estimated via BTYD models.

In Step 4, the planned enrichments of leads with the identified additional decision-critical insights were implemented technically. For example, to enrich the leads with key customer metrics estimated via the BTYD model, we conducted a complete data-mining process (Shearer et al., 2000). In particular, we built probabilistic models for estimating customers' future expected transactions, the probability of being alive (not churned), and estimates of future customer lifetime values (Fader and Hardie 2009; Platzer and Reutterer 2016). Probabilistic models of the family of BTYD models come from the field of marketing research and are particularly suitable for our setting, as they require comparatively little individual-level data (Rossi and Allenby 2003; Abe 2008; Van De Schoot et al. 2015; Platzer and Reutterer 2016).

In OEM's non-contractual market setting, technical managers with large fleets are handled by key-account managers. However, there are also many technical managers with small to medium-sized fleets that produce only small amounts of individual-level (i.e., ship-level) sales data. This can affect the predictive performance of machine-learning algorithms (Shaikhina and Khovanova 2017). On the other hand, BTYD models (Fader and Hardie 2009) like the Pareto/GGG (Platzer and Reutterer 2016) apply hierarchical modeling, which allows group-level information about the selected cohort of ships to be used when individual-level data is sparse (Efron and Morris, 1977).

At the start of the BTYD modeling procedure, the required input data had to be defined and extracted from OEM's data warehouse. The comparatively simple format of the required raw-data input is another strength of BTYD models, as they usually require only a transaction log of orders as instances, along with the order date, its value, and a unique identifier of the ship for which the order was placed. The transaction log is then transferred into the programming environment R, where the data is further processed and transformed into an aggregated higher-level format where the instances are ships and the variables are, for example, the number of transactions, the date of the first transaction, the logarithms of the timing between transactions, and the sum of the transaction values. Then, we estimated the parameters and hyperparameters of the selected Pareto/GGG model using a Markov-Chain-Monte-Carlo approach (MCMC; Platzer and Reutterer 2016). Based on this, future transactions can be predicted in a Bayesian way by drawing from the posterior distribution. Also, future probabilities that ships are active and alive can be calculated easily. Eventually, with an additional probabilistic model for monetary value (Fader and Hardie 2013), future CLVs can be calculated (Figure 7).

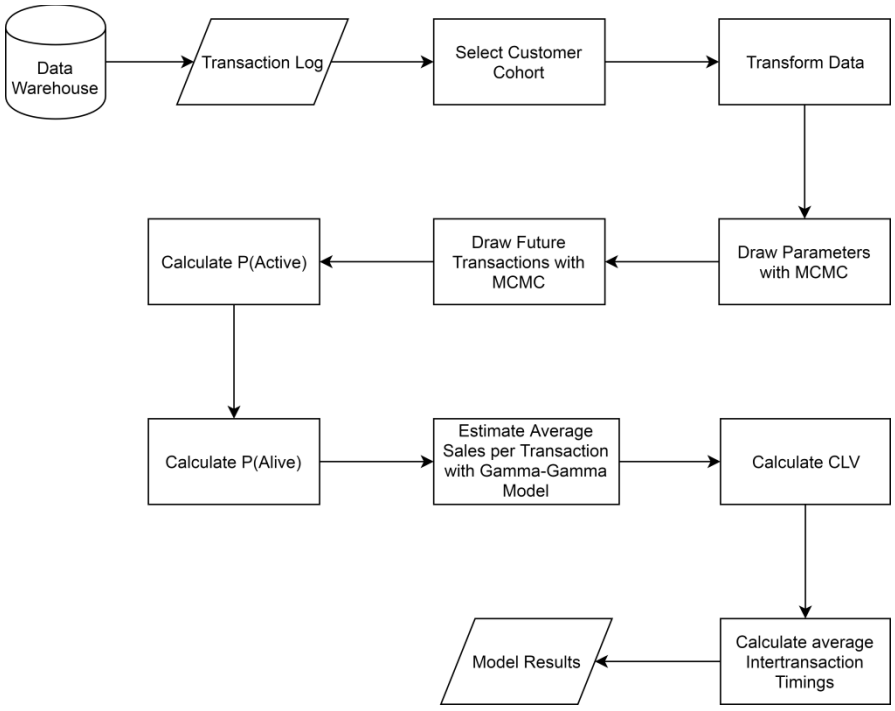

Figure 7. BTYD process (technical manager case)

We used around 500,000 orders as input data to predict the number of transactions one year ahead for around 24,000 ships. When we predicted the number of future expected transactions using the Pareto/GGG model that incorporates regularity parameters, the overall accuracy (predictions compared to test data) was 93 percent for predictions of the whole customer base. This reflects a high level of accuracy. On an individual level, the mean absolute error (MAE) was used as a performance metric. In the case of the Pareto/GGG, the MAE was 1.2.

When we applied the hierarchical Bayes version of the Pareto/NBD model, the accuracy of total predicted future transactions was 78 percent, with a MAE of 1.4 (Table 1).

Model	**Actuals / Predictions**	**MAE**
Pareto/GGG	93%	1.2
Pareto/NBD (HB)	78%	1.4

Table 3. Predictive Performance of BTYD Model

In **Step 5**, the leads were joined with the additional decision-critical insights in an SQL database, based on which spreadsheets were created and stored so they can be associated with a lead. Another important task was to specify to which sales employees a specific lead should be assigned. We reached out to the potential users in OEM's regional sales companies to get the required information. However, after the beta version was implemented, users became active themselves in specifying the correct assignees.

Step 6 supplied us with feedback regarding how to improve how leads are assigned and how additional decision-critical insights can be used to improve decision-making and action-taking. This step shows that the artifact encourages the use of feedback loops throughout the lead generation process. Moreover, the degree of uncertainty with which sales employees as decision-makers were confronted can be assessed as medium to low, as they had access to descriptive insights about the lead events and additional descriptive and predictive insights that together created the basis for decision making and action-taking. For instance, sales employees had insights into recent changes in technical management and the past sales history of customers so that they could answer the "what is the lead?" question (descriptive). Based on the predictive insights regarding future transactions and churn probabilities, they could also answer the "how will the customer relation probably be?" question (predictive). This combination of insights gave the sales employees at least partial answers to the "how to approach the lead?" question (prescriptive). They could, for instance, determine that a ship with a recent change in technical management with which OEM had a good customer relationship and that had a high predicted future CLV should be contacted immediately to avoid losing a valuable business relationship. In other cases, it can be sufficient for the sales employees to use the generated insights for adjusting the general way in which they approach a customer, e.g., adapting their marketing efforts based on the generated customer insights.

In Step 7, change management measures were formalized loosely following Kotter's (1995) eight steps:

1. **Establish a sense of urgency:** The data-driven lead-generation artifact was presented in several business meetings, and emails targeted to all assignees of leads were sent to create awareness and inform them about the sometimes-imperfect use of the CRM systems as the preferred sales tool.
2. **Form a powerful guiding coalition:** By contacting and involving the heads of sales of the local sales companies that were involved in the project and closely involving the responsible business managers, a powerful guiding coalition was built.
3. **Create a vision:** Together with a general digitization and business transformation initiative, the project drew on the vision of proactive 360-degree sales services with a customer-centric focus.
4. **Communicate the vision:** The overall vision of the digitization and business transformation initiative was continually communicated via managers, the intranet, and other internal communication channels.
5. **Empower others to act on the vision:** The department in which the ADR team was located organized periodic meetings in which stakeholders related to analytics could share knowledge and best practices. Here, the artifact and its various iterations and sub-steps were presented several times.
6. **Plan for and create short-term wins:** From the beginning on, the plan was to create quick wins by focusing first on a few local sales companies that are well connected to the department. This was done to create initial success stories and communicate them via the intranet in order to create awareness and support in the organization.
7. **Consolidate improvements and producing still more change:** Because of its iterative nature, the versions of the technical manager change lead pipeline were consolidated and improved several times as a result of the feedback loops.
8. **Institutionalize new approaches:** By showcasing qualitative and quantitative success measures, the technical manager change lead pipeline could be institutionalized and expanded to a broader target group.

4.4.3 Evaluation of the artifact

The resulting artifact and sub-artifacts were evaluated by the involved practitioners and their users throughout the entire design process. The evaluations were based on observational field notes, meeting notes, internal documents, informal interviews (especially collegial dialogue and joint problem-solving), and readily available data like usage reports on several decision-support systems and workshop outputs. We involved stakeholders and potential users continually to evaluate changes in the artifact. In particular, we were in close contact with the aftersales data analytics manager, the application manager of the CRM system, the team lead for analytics, and a selection of sales managers and sales professionals from OEM's sales companies, as potential users of the artifact. For example, we conducted at first informal interviews with potential users, then we refined the artifact based on the feedback, and eventually, we presented the refined artifact to a management audience. In short, we followed the ADR principles of reciprocal shaping, mutually influential roles, and authentic and concurrent evaluation, as Sein et al. (2011) proposed.

So far, around 650 leads have been generated in the CRM system directly from an instantiation of the data-driven lead-generation artifact. The initiative also inspired other data-driven lead campaigns that have generated another 2,000 leads. While in the beginning, just a handful of leads for one particular sales company were created, the pipeline's scope was quickly broadened to cover more than ten countries and their corresponding sales companies. This provides a clear indication that the aftersales organization sees practical value in the artifact. It is still too early to present metrics on the revenue generated through the artifact, as it is not linked directly to the quotation and ordering processes in the ERP systems. However, the feedback from the users has been positive, and the ADR team has received several suggestions for expanding the artifact to other campaigns beyond that of changes in technical management, such as to ship breakdowns and the sea trials of newly built ships. The feedback from the application manager of the CRM system has also been positive:

> *It's very interesting to see what scientific theories applied to our data sources can be used for. It has been important for us to include some of the receivers/end-users of the data-driven leads in the process to make it tangible for them and to gain from their real-life expertise and not end up*

with a bunch of leads that only looked promising on paper. Having their stamp of approval is the first step toward a more proactive sales process and creating additional value. The data-driven leads will be an addition to their work and will save them some time when they are looking for new leads in the market, as these leads come out of the box that is our CRM system.

The artifact has also fulfilled its objective of enabling proactive sales processes, as sales representatives can take immediate action and contact customers based on the leads without having to wait for an inquiry from the customer side. As the application manager of the CRM system explained:

We have to search for leads wherever we can, and using the data sources available is a natural next step in a more proactive sales approach. It's important that we set up an automated process around it and analyze the outcome of the data-driven leads to optimize the process over time.

One obstacle we observed is that the users of the CRM system have not yet fully adapted their work practices to the system's new capabilities. For example, users do not always document their work correctly in the system (e.g., setting a lead as qualified after being in contact with a customer).

The operational BTYD models to predict future customer behavior have been presented to managers and sales professionals on several occasions. The managers' evaluation was positive, and one business unit was interested in applying a similar approach to their particular business case. However, the feedback from sales professionals was mixed, as some saw the approach as too advanced, considering that it predicts future customer behavior while some of the current sales processes do not even use descriptive information. Nevertheless, the sub-artifacts, such as the BTYD models, have been seen as a positive outcome that, as an operational approach to analyzing the customer base, is applicable to many use cases and possible lead pipelines.

The project's main design objective was to design a DDD artifact that helps to create a new data-driven and proactive lead generation process in the CRM system. By applying an instantiation of the data-driven lead-generation artifact in the form of the technical management change pipeline, we created a new process for lead generation at OEM. Moreover, by integrating the

artifact with and framing it in DDD theory, we fulfilled the objective of creating a data-driven process. The initial design principle of embeddedness was incorporated into the artifact by embedding the DDD elements into a process of stakeholder involvement and change management. The initial design principle of partial automation was incorporated by including a stage of generating leads and assigning them to sales professionals in the CRM system automatically. Moreover, we automated the process of detecting lead events as well as the lead selection, and the enrichment of leads with additional decision-critical insights. Nevertheless, transferring leads to the right assignee still needed some degree of human involvement, due to organizational constraints (e.g., admin rights). The initial design principle of proactivity is incorporated in the artifact by supporting sales professionals with prescriptive insights about how to take action.

4.4.4 Growth of design theory

Over the cycles of the ADR project, we abstracted from the original problem and the original solution instance (i.e., the technical manager change pipeline) toward a more generic data-driven lead-generation artifact. Based on our experiences while designing and implementing these solutions, we learned why the artifacts we designed are effective solutions to the problems we encountered in the field. Following the ADR methodology, we formalized these lessons into design principles. (See Thiess and Müller 2018, for a more detailed presentation) We started out with initial design principles of being data-driven, proactivity, partial automation, and embeddedness, which were informed by the selected theory and the problem diagnoses. Throughout the building, intervention, and evaluation iterations, other nascent design principles emerged and subsumed the initial design principles. We formalized these generic and somewhat latent design principles by following the template Kruse et al. (2015) introduced, according to which the final design principles reflect material properties, enacted affordances, and constraints.

DP1: Theory-driven modeling – Given a lack of proof-of-concept, use theory-based models instead of data-driven machine learning algorithms to achieve concrete results.

Data-driven machine learning algorithms like gradient boosted trees and neural networks have proved their usefulness in work with large, high-dimensional datasets. They have shown their

superior performance compared to often more theory-based applications of models like logistic regression. Therefore, we first constructed an artifact based on a data-driven machine learning algorithm. However, it became apparent that the algorithms could not find meaningful relationships among the variables in the complex dataset, which led to overfitting and low performance of the model on unseen data. As a result, we could not implement our artifact in this form, leaving us without a concrete solution to the diagnosed field problems. Moreover, we lacked a theoretical foundation to guide our modeling, as our particular data-modeling application had not been made before, nor did an explorative data analysis or the consultation of domain experts reveal associations that could have enabled us to formulate an initial theoretical model.

Therefore, we turned to the marketing literature to search for alternative approaches to predicting customers' future purchasing behavior based on customer transaction data. Hierarchical Bayes models of the BTYD type satisfied our requirements for a modeling approach, as they were developed for the problem domain of non-contractual market settings and facilitate predictions of customer purchases, allow individual-level parameter estimations from group-level data, and require little data. Moreover, BTYD models are based on sound behavioral theory and, because of the possibility of using informative priors, do not require large amounts of data to produce good predictive performance (Van De Schoot et al. 2015). In the end, choosing a theoretically grounded modeling approach enabled us to create concrete results in the form of solutions to the field problem.

DP2: Comprehensibility – Limit models' complexity to gain support from managers.

In addition to predictive accuracy, comprehensibility heavily influences user acceptance of any decision-support system (Gregor and Benbasat 1999). We started out using machine learning to classify leads for aftersales from transactional customer data. This approach involved data-driven machine learning algorithms and a complex, highly dimensional dataset. As a result, even though the stakeholders' analytical background appeared to be strong, they had difficulty comprehending how the black-boxed algorithms processed the data to generate meaningful

insights. In contrast, the BTYD models that we used later in the project seemed to be easier to comprehend even though they are mathematically complex. An explanation for this observation may be that the BTYD models require only three pieces of information about each customer as input: their recency (i.e., the time of the last transaction), frequency (i.e., the number of transactions), and the monetary value of the transaction (for calculating CLVs). This information, which can be provided in the form of an event log of purchase transactions for each customer, was in line with the experience and intuition of the involved domain experts, minimizing the gap between the predictive model and managers' mental models.

DP3: Domain Knowledge – Incorporate domain knowledge into the data-driven decision-making process to encourage acceptance by managers.

Machine learning models can detect associations between variables related to customers' purchasing behavior from historical transactional data. However, human experts have developed expertise through years of experience in marketing and selling services to customers, domain knowledge that tends to be implicit and heuristic in nature (e.g., best practices, rules of thumb). This implicit knowledge is difficult to formalize but can hold valuable information for predicting future customer behavior. In our project, we included the knowledge of domain experts via business rules that capture an experts' experience and intuition regarding the context of the field problem. For instance, we interviewed domain-specific experts at OEM to collect data regarding the types of life-cycle events that constitute a demand for spare parts and service. In addition, the BTYD models are based on Bayesian theory, which allows us to incorporate beliefs about the relationship between input and output variables (e.g., how important the recency or frequency of past transactions are in predicting future customer behavior) in the form of informative prior distributions. Eliciting and incorporating this expert knowledge into the artifact increased the user's level of participation and influence on the final design, a key success factor in ensuring acceptance of the final artifact (Hollander et al. 1973).

DP4: Actionability – Provide actionable insights instead of quantitative reports to increase use by decision-makers.

Even if a decision support system produces highly accurate decisions and wide acceptance, it is not a given that end-users will follow those decision proposals and take action. Many organizations fail to take appropriate actions based on the generated insights because of the artifacts' focus on descriptive information and lack of prescriptive theory (LaValle et al. 2011). The departmental practitioners observed that analytical applications that are not based on a business process are used less often than those that are. Based on this early feedback, we decided to push the leads generated by our artifact directly into sales employees' daily newsfeed inside the CRM system instead of building extra reports or dashboards that the representatives would have to pull. The process was designed so every lead is created as a separate item and accompanied with additional information regarding what to do in the form of descriptive and predictive insights, but also via attaching clearly formulated prescriptive instructions with regard to what actions a sales employee should take, e.g., contacting the customer. Moreover, based on the meetings with the regional sales organizations, we decided to enrich the leads with additional ship and customer transaction information, so the sales employees have all the information they need for their regular lead follow-ups at their fingertips. This approach "makes it harder for decision makers to avoid using analytics—which is usually a good thing" (Davenport 2013).

4.5 Key Lessons

Our experience in applying ADR in a real organizational setting gave us the chance to be directly involved in the organization's operational processes. It provided us with an insider perspective that enabled us to collect and process large amounts of empirical data in the form of, for example, field notes, informal interviews, internal documents, and readily available data. Because of ADR's emphasis on authentic rather than controlled settings, we were able to inform our artifact design with empirical findings, creating a dynamic design process in which we could shape the artifact at high speed based on quickly executed cycles of building new features, demonstrating the features to practitioners, and evaluating their responses. However, we found it particularly challenging to report on our experiences during our design journey using the traditional structure for academic publications. For instance, if we wanted to report on every new design iteration, we would have had to describe more than a hundred micro-iterations.

Therefore, we decided to report on only major changes to the artifact and to group the micro-iterations that went into such major changes under one umbrella iteration.

Moreover, we found that it is helpful to distinguish initial theory-informed design principles from design principles that emerged throughout the design iterations because they represent two approaches to theory generation and so two research contributions. Theory-informed design principles that are tested via an artifact in a field setting represent a deductive approach and, thus, a deductive research contribution, while design principles that emerge out of the design process without being informed by theory represent an inductive approach, and, thus, an inductive research contribution. As a third research contribution and an attempt to generalize our context-specific findings, we found it helpful to relate and abstract both kinds of design principles back to theory when formalizing them in the last ADR stage, which represents an abductive theorizing approach.

In our next ADR application, we hope to use the notion of affordances more stringently as an analytical tool, as doing so will help to articulate the relationships among material properties, constraints, and intended effects in terms of user behavior (affordances; Seidel et al. 2018). We hope to use the ethnographic method of shadowing that is popular in interaction design to determine how material properties enact affordances and how those affordances may differ from those that are intended.

5 Paper V

Design Principles for Explainable Sales Win-Propensity Prediction Systems

By Tiemo Thiess,

Oliver Müller,

and

Lorenzo Tonelli

Abstract

MAN Energy Solutions, one of the largest ship engine manufacturers in the world, is looking into further improving its hit rate of through-life engineering services and spare parts quotations. We help to solve this relevant field problem by building a novel machine learning based sales win-propensity prediction system that utilizes the lightGBM algorithm, SHapley Additive exPlanations, and a second layer conditional probability model of quotation age. Moreover, we build an implementation method for the broader class of such systems and extend the scientific literature on explainable machine learning by abductively developing and instantiating the design principles (DPs) of local contrastive explainability, global explainability, selective visualization, causality, confirmatory nudging, and accountability in a sales win-propensity system.

Keywords: Machine Learning, Explainability, Sales, Maritime Industry, ADR

5.1 Introduction

In the last years, shipbuilders and original equipment manufacturers (OEM) in the maritime industry have suffered from a significant drop in the demand for new-building of vessels and engines (Danish Ship Finance 2018). An ongoing oversupply of tankers and containerships in the market caused this drop. OEMs are especially challenged to rethink their traditional business models and to shift the focus in product lifecycle management from the product development phase (beginning-of-life) to the product usage phase (middle-of-life). In the approximately 15-20 years lasting usage phase of main engines, OEMs can generate earnings via spare parts sales and through-life engineering services (TES), such as maintenance, repair, and overhaul. For OEMs, the product usage phase of their installed equipment determines the aftersales market.

In this context, MAN Energy Solutions, one of the largest ship engine manufacturers in the world with high market shares in the tanker and container vessel segments and approximately 15.000 employees in over 100 destinations around the world, is looking into further improving its hit rate[4] of through-life engineering services and spare parts quotations. Following the dual mission of IS, we help to solve this relevant field problem by building a novel sales quotation win-propensity[5] prediction system, while extending the scientific literature (Benbasat and Zmud 2006) on explainable machine learning by abductively developing design principles (DPs) based on a sound literature review and an authentic and concurrent evaluation of the action design research (ADR) process (Sein et al. 2011).

Win-propensity estimation is an important aspect of assessing overall sales performance (Monat 2011). Despite its importance, research on sales win-propensity estimation models is scarce (Yan, Zhang, et al. 2015). In large firms such as MAN Energy Solutions, sales professionals sometimes have to deal with many open sales opportunities and quotations. To structure their work and to enable an approximate forecast of the win-propensity, sales professionals use CRM

[4] At MAN hit rate is essentially calculated as orders euro / quotations euro (ex-post)

[5] Win-propensity is the hit rate expressed as a probability for a particular quotation (ex-ante)

systems that enable them to assign win-propensity scores or hot-warm-cold labels manually as an outcome of an often more or less subjective judgment (Xu et al. 2017). Such subjective judgments are prone to cognitive biases (Tversky and Kahneman 1974), such as being overly confident and thus estimating too high win-propensity scores (Bohanec, Robnik-Šikonja, et al. 2017a). Moreover, they can be biased due to organizational structures, politics, and socio-cultural phenomena, for example, when the management expects a positive forecast for the current sales pipeline (Yan, Zhang, et al. 2015). Data-driven sales win-propensity estimation methods, on the other hand, can support resource management (D'Haen and Van den Poel 2013), increase efficiency, and generate explanatory insights about the sales process and its drivers (Bohanec, Kljajić Borštnar, et al. 2017).

Overall, in this paper, we make four scientific contributions. First, we push the state-of-the-art in sales propensity modeling by developing an approach combining ensemble machine learning techniques (esp. lightGBM) to robustly model non-linear relationships and interaction effects with a conditional probability model accounting for quotation age (Sections 5.4.1 and 5.4.2). Second, we demonstrate how methods for the human-interpretable explanation of black-box machine learning models (esp. SHapley Additive exPlanations) can be applied to improve the acceptance of predictions by users and managers, and how they help data scientists to improve model quality (Section 5.4.3). Third, we go beyond the pragmatic design of a single prototype and propose a method for the organizational implementation of the proposed approach in complex real-life settings (Section 5.4.4). Fourth, we formalize the learnings from this 1.5 years lasting action design research project as design principles for explainable aftersales win-propensity prediction systems (Section 5.5).

5.2 Explainable Machine Learning

Machine learning and data science have the ultimate goal of supporting decision making. Common sense tells us that one should only implement good decisions. But what are good decisions? Sharma et al. (Sharma et al. 2014) present two characteristics of good decisions: quality and acceptance. The quality criterion is concerned with whether a decision is able to reach its stated goals. The other criterion refers to whether a decision is accepted by its stakeholders, es-

pecially those responsible for successfully implementing it (Drucker 1967; Hollander et al. 1973; Sutanto et al. 2009). Hollander et al. (1973) argue that how much stakeholders participate in the decision making process and, thus, influence the final decision, significantly impacts its acceptance and the chances of successful implementation.

Also, Kayande et al. (2009) suggest that a lack of understanding of a machine learning model can lead to a refusal of acceptance and, consequently, usage by end-users, despite the fact that the model might improve decision quality. They further elaborate on this idea by proposing a three-gap framework that conceptualizes how human-interpretable explanations can be used to improve the acceptance and performance of decision support systems (DSS). In particular, they relate three different concepts, namely, the manager´s mental model, the DSS, and the true model (reality) via three distinct bi-directional gaps. The first gap, between the manager´s mental model and the DSS, can lead to reduced model acceptance when widened and improved model acceptance when narrowed. The second gap, between the DSS and the true model (reality), affects the performance of a DSS negatively when widened and positively when narrowed. The third gap, between the manager´s mental model and the true model (reality), affects the manager´s decision making performance negatively when widened and positively when narrowed.

Gregor and Benbasat (1999) give arguments for why users need explanations when working with intelligent systems such as machine learning systems, namely, to solve specific problems by using the system, to learn from the system and its outputs, and to understand why anomalies have come to be. Moreover, they argue that explanations can lead to an improvement in terms of performance, learning, and the overall perception of a system. However, they also note that in order to enable such improvements, explanations should be context-specific rather than too generic and should not demand too much effort from users and, thus, if possible, be automated. Finally, they stress the importance of justificatory knowledge, which can lead to a deeper understanding by grounding, for instance, a prediction in sound causal theory.

Martens and Provost (2014) have extended both the work of Kayande et al. and Gregor and Benbasat. They criticize the three-gap framework of Kayande et al. because it assumes that DSS are always superior to a manager's mental model in terms of decision quality (alignment with reality). Instead, they argue that DSS can be wrong too, for example, because of biases introduced during the model building process or overfitting a model to training data. In consequence, they extend the three-gap framework by adding a feedback loop for situations in which a manager´s mental model is closer to the true model (reality) than the DSS. The objective of this feedback loop is to improve the DSS by bringing it closer to the manager´s mental model. Moreover, they add the three different roles of developers, managers, and customers to aid understanding of how the explanatory needs of the roles differ.

Furthermore, Martens and Provost (2014) extend the above-outlined arguments of Gregor and Benbasat by distinguishing between (1) explanations that lead to improved system acceptance by supporting the user in getting a causal understanding of the general real-world mechanisms that the system builds upon and (2) explanations that lead to improved acceptance by supporting the user in understanding how the particular system works. They further subdivide explanations of type (2) into (a) global explanations of how the overall model behaves and (b) local explanations of how it behaved in a particular instance. Such types of explanations, they argue, can lead to improved acceptance but also an improved model, which, again, can improve model acceptance but also aid in making sense of the model´s underlying causal mechanisms (reality).

5.3 Methodology

We followed an action design research (ADR) process inspired by Sein et al. (2011), in which we started by analyzing and formulating the field problem of aftersales hit rate improvement at MAN Energy Solutions. Next, we designed initial artifacts of the class of explainable win-propensity prediction systems. Throughout many iterations of building, intervention, and evaluation, the artifacts were further shaped and refined by the design team, but also by the specific context of the maritime industry, until they reached their current state. Finally, we formalized abstracted learnings as design principles for explainable win-propensity prediction systems. During the whole process, we collected rich empirical data in the form of observation notes of

our encounters at MAN Energy Solutions (see Table 1 for an overview). To collect the data, we used a form of design ethnography (Baskerville and Myers 2015), in which one does not only study others and their behavior, but also oneself and one's artifacts, and how they interact as interventions with their environment.

Table 1: Project-related encounters at MAN Energy Solutions

Meeting Type	*Participants*	*#*	*h*	*Total (h)*
Development Meeting	Business Analyst, Junior Data Analyst, Data Management Specialist, Researcher	20	2	40
Stakeholder Presentation	Business Analyst, Department Manager, Sen. Strategy Manager, Strategy Manager, Researcher, Pricing Analyst	4	1	4
Sprints	Researcher, Business Analyst	18	4	72

5.4 An Explainable After-Sales Win-Propensity Prediction System

At MAN Energy Solutions, we built a system for win-propensity scoring that is integrated into the existing IT infrastructure (see Figure 1). In the spirit of ADR, this system constitutes the main practical contribution of our work. The core of the system is a lightGBM model (Ke et al. 2017) that produces base win-propensity probabilities for sales quotations and a second-level conditional probability model that accounts for the decaying effect of quotation age on the base win-propensity probabilities. Moreover, we train a separate explanatory SHAP (Shapley Additive exPlanations) model to open up the lightGBM black-box model by generating human-interpretable explanations of global and local (individual predictions) feature importance (Scott M. Lundberg et al. 2018). The model training part of the system executes over the weekend, while the prediction part of the system executes daily. Both parts work fully automated.

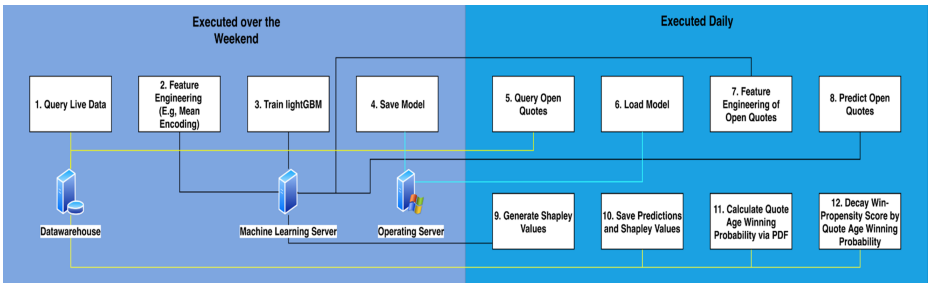

Figure 1. Implemented back-end process

5.4.1 lightGBM-based Win-Propensity Prediction Model

lightGBM (Ke et al. 2017) is an advanced implementation of the boosting algorithm (Freund and Schapire 1996) that utilizes gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS excludes instances with small gradients (residuals) from the data and focusses on instances with larger gradients to compute information gain. By this, lightGBM is faster than other implementations of boosting. With EFB, lightGBM can further improve performance by reducing the number of features via bundling variables that are mutually exclusive together. We chose lightGBM mostly due to its performance properties. In our case, the lightGBM algorithm was by far the most efficient tree ensemble method when trained on our data of up to 3 million records of quote positions and 15 carefully selected features (see Table 2). Through cross-validation, we get an average accuracy of 76% and an AUC (area under the receiver operator curve) of 0.74; meaning that there is a chance of 74% that the model can successfully distinguish between a randomly selected won quotation and a randomly selected lost quotation. Compared to a model with no separability power (AUC of 0.5), our model provides a lift of 24%. Furthermore, as the model calculates win-propensity probabilities with an average Brier score of 0.20 and not just binary labels, its outputs can be used by sales professionals directly to evaluate and prioritize sales quotations.

Table 2. Example of features used in the model

Feature	Equip. in Plant (id)	Material (id)	Discount (percent)	List Price (euro)	Processing Time (days)
Encoding	Mean	Mean	Numeric	Numeric	Numeric

5.4.2 Second-level Conditional Probability Model for Quotation Age

During the implementation of the system, we faced the challenge of incorporating the time-dependent decay of win-propensity probabilities into the lightGBM model (in other words: the older a quotation gets, the less likely it is that it will be transformed into an order). The technical problem was that for all non-hit training records (i.e., rejected quotations that were never transformed into orders), we lack the reference (order date) to calculate the difference in days between quotation creation and order date.

To overcome this challenge, we developed a two-layer modeling approach. The first step in this approach is to estimate the probability density function (PDF) of the win-propensity score for quotation age. For this, we only use the subset of won-quotations that have an order date. We first calculate the difference between quotation creation and order date (quotation age) and then calculate the frequency of won-quotations and group them by quotation age. As we have access to a large amount of data and quotation age is a continuous random variable, we chose to estimate the PDF via a histogram, which as a non-parametric estimation method is suitable in this case (Izenman 1991). From the PDF, we can draw probabilities for each quotation age. Eventually, we calculate the time-decayed win-propensity as a conditional probability and integrate it into the user interface (Equation 1 and Figure 2):

$$P(Propensity \mid Age) = \frac{P(Propensity\ and\ Age)}{P(Propensity)} \quad (1)$$

sales_order	HR pred	HR pred (time adjusted)	edit	Customer (number & name)	Motor Manager (name & number)	Vessel
	95 %	53 %				
	86 %	53 %				
	93 %	52 %				

Figure 2. Quotation view of win-propensity prediction (HR pred) and time-adjusted prediction (blurred for confidentiality reasons)

5.4.3 SHAP Model

SHAP (Scott M. Lundberg et al. 2018) is grounded in the game-theoretical concept of Shapley values. If one imagines that players are collaborating in a team (coalition) to win a game, then Shapley values are the marginal contribution of a player´s performance to the overall success of

the team. Based on Shapley values, all players could be paid fairly by their clubs according to their contribution to winning the game.

Machine learning researchers adapted this idea and developed algorithms for local-level machine learning model explanations (per prediction; Scott M. Lundberg et al. 2018; Lundberg and S.-I. Lee 2017; Lundberg and S. I. Lee 2017; Štrumbelj and Kononenko 2011, 2014). The idea here is that the prediction, in our case predicting win-propensity, is the game, and the feature values are the players. Thus, if we can calculate the marginal contribution of each feature value, we have a consistent method of feature importance that is superior to standard feature importance methods such as gain (in terms of Gini index) or splitcount. Moreover, compared to the local interpretable model-agnostic explanations algorithm (LIME) (Ribeiro et al. 2016c, 2016b), SHAP is more interpretable, since its explanation values add up to the model output. Also, SHAP-based global feature importance allows visualizing non-monotonic relationships (bi-directional; see Figures 3 and 4).

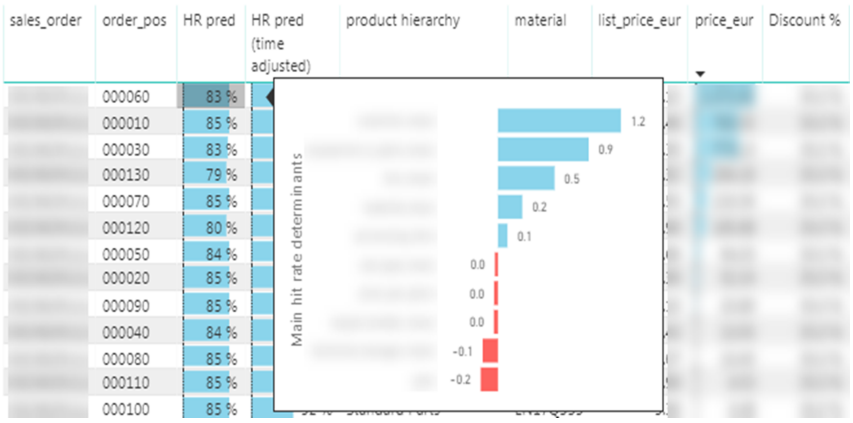

Figure 3. Local instance-level SHAP explanation (blurred for confidentiality reasons)

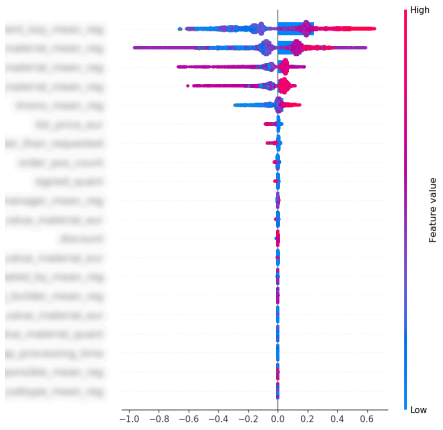

Figure 4. Explanations showing both magnitudes (X-axis) and non-monotonic value impact (color code; the figure is blurred for confidentiality reasons)

5.4.4 Implementation Method

Next to the core system consisting of back-end and user interface, we designed an implementation method for our system that was, despite its initial theoretical grounding, developed abductively based on the learnings of the different ADR building, intervention, and evaluation cycles (see Figure 5). One notable highlight of the method is the utilization of domain knowledge to develop hypotheses of drivers of hit rate that, in the following, are tested via SHAP in an explanatory analysis. Furthermore, we add steps of regular stakeholder presentations as well as the utilization of UI mockups for those presentations. Also, we distinguish between model building and evaluation in experimental lab situations and in a more naturalistic production environment, which, in our experience from the case, can give different results and, therefore, valuable insights into the modeling and implementation process. Moreover, we added steps of deployment in production UI, change management, and live monitoring to account for the fact that the data generating processes that our models rely on may change.

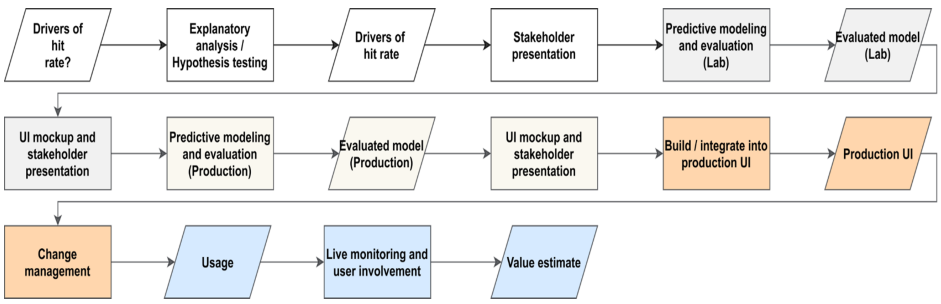

Figure 5. Implementation method

5.5 Design Principles

The following design principles abstract from the concrete artifacts described in Section 4 and capture general prescriptive knowledge that should enable others to build explainable machine learning systems, in particular, win-propensity prediction systems. In the spirit of ADR, these design principles constitute the main scientific contribution of our work.

5.5.1 Local-contrastive Explainability: Present model explanations to users on an instance-level to support contrastive explanation processes

Due to the complexity of the aftersales market in the shipbuilding industry and its industry-specific challenges, such as intransparent owner structures, bulk orders, and product heterogeneity, sales professionals at MAN rely on implicit domain knowledge and collaborate closely with other domain experts (e.g., engineers). Not surprisingly, we found that these sales professionals often do not trust the predictions of a machine learning model. While the sales professionals are not necessarily interested in fully understanding how the machine learning model has generated a score, they still want to know why the model predicts a specific win-propensity score for a specific quotation. They may ask: "Why has quotation X a win-propensity score of 0.9 and not 0.2?" According to Lipton (1990) and Miller (2019), answering such why-questions requires the explainee to contrast the observed event (score of 0.9) with an imagined counterfactual event (score of 0.2) to abductively infer the most plausible explanation of the observed event (score of 0.9; Harman 1965). We support this cognitive process of abductive reasoning by displaying Shapley values alongside main variables such as discount % or material in our user interface (Figure 3). In the abduction process, the Shapley and variable values function as candidate hypotheses of causes for the observed event (score of 0.9), which users can compare with their mental model to assess their plausibility and eventually answer the contrastive question: "Why has quotation X a win-propensity score of 0.9 and not 0.2?"

5.5.2 Selective Visualization: For local explanations, visualize only the top contributing features to reduce explanation complexity

Our front-end shows a report of open quotations along with the predicted win-propensity scores and a time-decayed version of it that accounts for quotation age. This report already con-

tains much information, and processing it puts a high cognitive load on users. Hence, we decided to not increase the information processing load further with our explanations. Instead, we wanted to limit the complexity of our instance-level explanations by providing on-demand visualizations of only the top-5 most important positive and negative features.

This empirically motivated design decision can be backed up with psychological theory. Psychological research suggests that human short-term memory can only recall 4-7 chunks of information at a time (Cowan 2001; Miller 1956). A chunk is the largest unit of information that human memory can represent. How the human brain creates these chunks depends on its prior knowledge. When confronted with familiar concepts, our brain can create larger chunks, and therefore, recall more information. Visualization supports this cognitive chunking process by grouping (or pre-chunking) information into symbolic representations so that one can display much larger amounts of information that otherwise could not have been recalled simultaneously (Larkin and Simon 1987).

5.5.3 Accountability: Schedule regular management presentations to increase data scientists´ need for justification

In our implementation method (Figure 5), we propose repeated stakeholder presentations to create and sustain organizational support. Committing to those presentations comes with the side effect of having to justify one´s approach and progress to the stakeholders. As a result, one can be made accountable for what one has done between the meetings.

Research from the field of psychology suggests that accountability, the need for justification of one´s viewpoints towards other people, lets decision-makers judge in more complex ways, rely less on prior beliefs, and be more evidence-based (and supposedly more data-driven). By this, accountability affects decision making in a debiasing way (Simonson et al. 1992; Philip E Tetlock 1985; Tetlock et al. 1989). For a developer (data scientist) that follows our method, this self-created accountability increases the need to understand how its machine learning model works. Thus, it motivates developers to align the gap between their mental model and the machine learning model (Martens and Provost 2014), which eventually can lead to improved model

quality. Moreover, having a solid understanding of a machine learning model is a pre-requisite for explaining it to others in a simple, but not simplistic, way.

5.5.4 Global Explainability: Explain the machine learning model to managers on a global level to increase acceptance, enable process accountability, and share outcome accountability

In our implementation method, we included a step of explanatory analysis/hypothesis testing, based on our experience that managers became much more engaged, contributed with domain knowledge, and seemed to be more positive towards the project, once we presented the results of our explanatory analysis. We presented not only findings concerning the drivers of hit rate but also how the different feature values on average affect the prediction of the model (global explanations).

While there is, as mentioned above, research that indicates a positive impact of accountability on decision making, there is also research suggesting negative forms of impact for some types of accountability (Lerner and Tetlock 1999; Simonson et al. 1992). In particular, Simonson and Staw (1992) suggest that outcome accountability triggers a mechanism by which accountable persons perceive an increased need to self-justify past behavior and decisions, which, in turn, leads to an escalation of commitment to such behaviors. Process accountability, on the other hand, leads to a more thorough alternative evaluation in decision making, but also a decrease of the need to self-justify past behavior, since one can justify behavior via a thoroughly evaluated and transparently reported process instead of exploiting or defending an eventual outcome only.

Based on our experience from the case, we argue that in machine learning projects, it is difficult for managers to comprehend the complex processes and mechanisms that underlie a system. In reaction to this, managers may tend to make developers (data scientists) outcome accountable. However, when faced with a task such as implementing a novel machine learning system, where the outcome uncertainty is high, outcome accountability increases the stress-level of data scientists (see Siegel-Jacobs and Yates 1996). The reason for this is that in high outcome-uncertainty situations, it is particularly challenging for evaluators to assess the effort-outcome

relation so that even when data science teams deliver high-quality work, the project can fail due to factors that are out of their control. In such situations, process accountability may be preferable to relieve some of the stress related to the low effort-outcome reliability (Wiseman and Gomez-Mejia 1998) and its negative consequences (Siegel-Jacobs and Yates 1996).

Nevertheless, it is hard to evaluate the quality of a process if it is not explainable. We experienced that presenting global explainability methods such as average SHAP feature importance (Figure 4) to managers, helps them to align their mental model with the machine learning model and the mental model of the data scientist. It allows evaluating whether a course of action (process) chosen by the developer makes sense or not, which eventually enables managers to make data scientists process accountable and, consequently, share the outcome accountability of machine learning projects and systems.

5.5.5 Confirmatory Nudging: Use language and representation devices that align well with users´ and managers´ mental models to increase acceptance of the machine learning model

In our system, we made sure that we use a vocabulary (esp. feature names) that is familiar to the stakeholders from the maritime industry, such as motor manager (customer), or equipment_in_plant (engine installed on a vessel) instead of non-speaking feature names such as x1, x2, x3 or features names that are uncommon in the given company and industry. Using such names when explaining the machine learning model helps to narrow the gap between a stakeholder´s mental model and the machine learning model, which, in turn, should increase its acceptance. Moreover, we made sure to present a working prototype early on (Figures 2, 3, and 4) by integrating the machine learning scrips and models into the existing infrastructure, which enabled us to demonstrate the model in an already familiar user interface.

Confirmation bias (Nickerson 1998) describes a tendency to favor information that aligns well with one´s prior beliefs (mental model). By adding domain-specific traits to the system, we exploit this cognitive bias to influence the behavior of users and managers in a predictable positive way (nudging; Thaler and Sunstein 2009).

5.5.6 Causality: Choose the machine learning model that aligns best with reality and design it as if it was an explanatory rather than a predictive model to increase model acceptance by users, managers, and developers

Shmueli [44, p. 293] discusses the differences between explanatory and predictive modeling, amongst others, based on the following two characteristics. (1) Causation-association: "In explanatory modeling *f* represents an underlying causal function, and *X* is assumed to cause *Y*. In predictive modeling *f* captures the association between *X* and *Y*". (2) Theory-data: "In explanatory modeling, *f* is carefully constructed based on *F* in a fashion that supports interpreting the estimated relationship between *X* and *Y* and testing the causal hypotheses. In predictive modeling, *f* is often constructed from the data". In our implementation method, we incorporated a step of reaching out to the business in order to develop hypotheses (low-level theory) of how the features relate to the target. In the next step, explanatory analysis, we are testing those hypotheses with accessible data. So instead of looking for associations only, a typical approach when the objective is mostly predictive, we start with developing a causal theory of how the features relate to the target variable, which is common when the objective is explanatory.

A more technical distinction between explanatory and predictive modeling objectives is the treatment of multicollinearity (Wheelwright et al. 1998, p. 288): "Multicollinearity is not a problem unless either (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to Y, without the influence of the other explanatory variables. Multicollinearity will not affect the ability of the model to predict." Also, SHAP (Figures 3 and 4) assumes independent features (Scott M. Lundberg et al. 2018). A violation of this assumption can bias the Shapley values for dependent (multi-collinear) variables since the algorithm cannot attribute the distinctive contribution of each feature to the prediction. In reaction to that, we treat multicollinearity like one would do when having purely explanatory objectives. First, we identify collinearity via correlation matrices and multicollinearity via variance inflation factor (VIF) analysis. Based on this and the developed causal model, we remove the collinear variables or merge them.

To summarize, we are designing the model to achieve both predictive and explanatory objectives. In our case, this comes with the benefit of increased explainability, while keeping the loss in predictive power neglectable.

5.6 Discussion and Conclusions

In this paper, we presented an explainable two-level win-propensity prediction system that utilizes the lightGBM algorithm (5.4.1), a conditional probability model for quotation age (5.4.2), SHAP explanations on both local and global levels (5.4.3), an interactive user interface (5.4.3), an implementation method (5.4.4), and abductively developed design principles (5.5).

To the best of our knowledge, our work is the only one that provides an implementation method and derives design principles based on learnings from designing and implementing a novel sales win-propensity prediction system in a real-world environment. Also, there is no other sales win-propensity approach for predicting the probability of converting a sales quotation into a sales order (hitting). Moreover, there is no specific machine learning approach for sales predictions in the maritime manufacturing industry.

Nevertheless, there are some approaches for predicting the win-propensity of sales leads or opportunities, which is a comparable sales conversion process that, however, happens earlier in the sales funnel. In this domain, researchers from IBM developed with OnTARGET a logistic regression model that predicts the propensity of customers to buy IBM´s products (Lawrence et al. 2010). Zan et al. (2014) applied a neural network-based approach. Yan et al. propose a win-propensity approach based on modeling the interaction of users with the sales support system as Hawkes Processes (Yan, Zhang, et al. 2015). Duncan and Elkan propose a pure probabilistic model (2015). Compared with these approaches, our approach is theoretically either superior in terms of predictive or explanatory power, and always superior in balancing predictive and explanatory power.

The approach by Bohanec et al. (2017b, 2017a) and Eitle and Buxmann (2019) are the only other approaches that come close in terms of theoretical predictive and theoretical explanatory power. They utilize with random forest and gradient boosting machines some of the empirically proven

best-performing prediction algorithms, that, however, are more resource-intensive when compared to lightGBM. Furthermore, they do not address the issue of time-decay in win-propensity scores that we approach with our second layer conditional probability model. Also, the explanation methods IME and EXPLAIN (Štrumbelj et al. 2009; Robnik-Šikonja and Kononenko 2008) used by Bohanec et al. and LIME (Ribeiro et al. 2016c) used by Eitle and Buxmann do not fulfill the criterion of explanation accuracy that SHAP fulfills (Scott M. Lundberg et al. 2018). None of the approaches explicitly deals with multi-collinearity, which potentially makes their approaches less aligned with reality, and due to this less suitable to align well with the mental models of users, managers, and also their own mental models, which in turn can lead to decreased trust, low acceptance, and low model quality (see Kayande et al. 2009; Martens and Provost 2014).

We build our approach during a 1.5 years lasting ADR project at MAN Energy Solutions. It means that the final shape of the system, the implementation method, and the design principles are not necessarily generalizable to other environments. However, they should be transferable to similar environments that are concerned with similar problems. While our system (Section 5.4) deals with challenges that should be transferable to many other B2B environments, our design principles are even further abstracted to the class of explainable machine learning, which should be transferable even to B2C environments.

In the future, we want to study further how the system, with its explanatory capabilities, affects the acceptance by both users and managers. Moreover, we want to compare the accuracy of the win-propensity predictions generated by the system with those generated by users. Furthermore, we plan to integrate the aggregated win-propensity predictions for a current sales pipeline as an operations-level forecast into a more general strategic forecasting algorithm.

6 Paper VI

Designing Causal Inference Systems for Value-based Spare Parts Pricing – An ADR Study at MAN Energy Solutions

By Tiemo Thiess

and

Oliver Müller

Abstract

In the wake of servitization and increased aftersales competition, original equipment manufacturers (OEMs) begin to change their pricing strategies from traditional cost-based to value-based pricing. As value-based pricing is much more individualized and data-driven, it becomes increasingly important to validate one's pricing hypotheses by estimating the causal effects of pricing interventions. Randomized controlled trials (RCTs) are conceptually the best method for making such causal inferences. However, RCTs are complicated, expensive, and often not feasible. MAN Energy Solutions was facing a similar challenge. In reaction to his, we conducted an action design research study (ADR) in which we designed and implemented a novel causal inference system for value-based spare parts pricing. Based on this, we formalize design princi-

ples for the broader class of such systems that emphasize the need for pre-aggregation when dealing with lumpy aftersales data, scalability when having to run numerous analyses on heterogeneous spare parts portfolios, and incorporating unaffectedness conditions that help to avoid spillover effects caused by often interdependent spare parts purchases. Also, they encourage analysts to take pre-intervention predictability into account when interpreting causal effects, to incorporate a manipulated treatment variable into the causal inference model, and to present the system output in interactive user interfaces to aid understanding and acceptance.

Keywords: Causal inference, value-based pricing, action design research, spare parts, industrial marketing.

6.1 Introduction

Original equipment manufacturers (OEMs) have, in the last years, started to focus on the usage phase of their products, instead of focusing on the product development phase only (Sundin 2009). In the product usage phase, they can generate earnings via spare parts sales and through-life engineering services like maintenance, repair, and overhaul (Cohen et al. 2006). This requires OEMs to develop new business models and customer-centric processes, a transformation that is enabled by new technology and data-driven approaches (Huang and Rust 2018; Rust and Huang 2014; Thiess et al. 2020; Thiess and Müller 2018). However, due to an intense competition of third-party companies that copy and sell non-original spare parts at low prices, the potential aftersales gains for OEMs are threatened (Gallagher et al. 2005). Even though pricing has the highest impact on earnings before taxes and interest rates (EBIT) (Hinterhuber 2004), OEMs struggle to realize its potentials, due to difficulties arising from vast portfolios of often thousands of different spare parts. In reaction to this, they employ undifferentiated cost-based pricing strategies (Gallagher et al. 2005). Instead, research suggests that OEMs should shift from cost-based to value-based pricing (Hinterhuber 2004, 2008; Hinterhuber and Liozu 2014; Wickboldt and Kliewer 2018) strategies. In a value-based pricing approach, one sets prices based on the value that materials provide to the customers, e.g., expressed in customers' willingness to pay. This approach incentivizes OEMs to innovate and develop their products according to their customers' needs and for maximal customer utility (Gallagher et al. 2005).

Nevertheless, implementing value-based pricing increases the complexity of the pricing function substantially, as one cannot use one-size-fits-all solutions. Instead, one needs to approach pricing much more individualized, which involves developing numerous, often data-driven (Andersson and Bengtsson 2013; Cullbrand and Levén 2012; Wickboldt and Kliewer 2018), pricing approaches based on various assumptions and hypotheses. However, testing such hypotheses requires more than uncontrolled before and after analyses (Goodacre 2015) or simple correlational approaches (Hernán et al. 2002). Instead, researchers generally perceive randomized controlled trials (RCTs) as the best method for making such causal inferences (Cartwright 2007; Cochrane 1972). However, in practice, RCTs are complicated, expensive, and often not feasible (Cartwright 2007; Varian 2016). Based on this, we formulate our main research question:

How to design causal inference systems that support value-based spare parts pricing decisions?

Here we also follow calls for more research on applications of artificial intelligence (AI) systems in B2B marketing contexts in general (Mora Cortez and Johnston 2017) and B2B pricing strategies in particular (Martínez-López and Casillas 2013).

To answer the research question, we conducted an action design research study (ADR) (Sein et al. 2011) at MAN Energy Solutions, one of the largest OEMs in the maritime industry and a world market leader for large-bore diesel engines. MAN Energy Solutions had recently started an initiative for implementing more value-based pricing strategies (Hinterhuber 2004). As a result of this, prices for more than 30.000 spare parts at one of the company's headquarters had been either de- or increased. Due to this, the company was facing the challenge of assessing and quantifying whether the interventions were successful, and the hypotheses behind the different value-based pricing initiatives could be validated. Historically, the company had used uncontrolled before and after studies in which one compares an outcome (e.g., sales volume) for a treated unit (material for which the price was changed), before and after an intervention (price change). While such approaches can be indicative, they are not suitable for estimating causal treatment effects (Goodacre 2015).

We identified this relevant field problem as a research opportunity to design and implement a novel system for value-based spare parts pricing support that is based on the state of the art of causal inference theory and addresses aftersales-specific challenges. Moreover, we go beyond the situated implementation of a system (Gregor and Hevner 2013) and abstract some of the principles that underly our design to the class broader class of such systems. Thus, we follow the dual mission of information systems research (IS) to create utility for practitioners while contributing to the scientific body of knowledge (Benbasat and Zmud 2003).

We proceed as follows. The next section provides the background on the causal inference approaches that informed our initial system design. Then, we discuss our overall research method. We then describe the system that we designed and implemented at MAN Energy Solutions. After that, we present and discuss a set of design principles. The final section discusses implications for research, practice, limitations, and concludes the paper.

6.2 Causal Inference on Observational Data

Causal inferences from observational data are challenging to make as one has to eliminate spurious correlations and rely heavily on the analyst's subject-matter knowledge when specifying causal models (Hernán et al. 2002).

Randomized controlled trials (RCTs) avoid such pitfalls when set-up ideally (Cochrane 1972) and are often called the gold standard of causal inference (Cartwright 2007). In an RCT, one randomly assigns subjects to different groups so that each group, on average, is more or less similar and, thus, comparable. Then, one manipulates one variable of interest (the treatment or intervention), while keeping everything else equal. This way, one can be confident that the treatment was the only cause of changes in the subjects. In such a situation, one can calculate treatment effects by taking the difference in means of the outcome variables for the treated and untreated control or placebo groups (e.g. Cochrane 1972).

However, setting up RCTs is complex and often not feasible (Schulz et al. 1995) as it is based on strict assumptions and requires random assignment to make inferences about how a treatment

affects a population of subjects (Cochrane 1972). In industry, however, one is usually more interested in how the treated subjects affected the firm (the experimenter; Varian 2016).

Donald Rubin introduced the potential outcome framework to causal inference (Rubin 1974). In this framework, one tries to compare what happened to a subject after receiving treatment (actually observed outcomes) with what would have happened had the subject not received treatment (the counterfactual outcomes).

Inspired by this, researchers have developed quasi-experimental methods (Cook et al. 1979). Such methods are based on observational data, but construct control groups from the data to estimate counterfactual outcomes. One of the most common approaches here is the difference-in-differences design (DID; Ashenfelter and Card 1985; Card 1990). DID treat time-series data of the observed outcomes for treatment and control groups as cross-sectional data and model them via standard ordinary least squares (OLS) regression by adding a dummy variable for the pre-intervention period (zero for all records) and the post-intervention period (one for treatment records and zero for controls). This way, one can estimate average treatment effects by comparing the difference between pre- and post-period outcomes of the treatment with the difference between pre- and post-period outcomes of the control groups. The idea here is that the control group represents the counterfactual situation that would have occurred had the treatment group not received treatment. While much more robust than uncontrolled before and after studies (Goodacre 2015), DID designs have some underlying assumptions that often do not fit reality (Abadie 2005). Most of all, they usually only consider single points in time before and after the intervention (dummy variable) without considering how effects develop throughout several post-intervention periods (Brodersen et al. 2015). Moreover, they assume independent and identically distributed (iid) random variables, which neglects the autocorrelation structure of time-series data and leads to biased estimates (Bertrand et al. 2004).

Newer approaches in the form of synthetic control methods (SCM) are more robust to biases. In such methods, one combines several potential control groups into one synthetic control group that, in the post-intervention period, represents the counterfactual outcome that one would

have expected had the treatment not occurred (Abadie et al. 2010; Abadie and Gardeazabal 2003). In the most common synthetic control approach as suggested by Abadie et al. (2010) and Abadie and Gardeazabal (2003), one constructs a synthetic control as the weighted average of all potential controls (adding up to 1) that minimizes the difference between pre-intervention treatment and control matching variables. This approach has limitations, as it requires access to explainable matching variables (e.g., age or gender of subjects) and does not allow for non-convex optimization approaches in determining weighted averages of control groups (Brodersen et al. 2015).

According to Hal Varian, the chief economist at Google, alternatively, one can define the estimation of counterfactuals as a prediction problem (Varian 2014, 2016): "[i]n this case, the counterfactual is the [prediction] of the outcome for the subject constructed using data from before the experiment. To implement this approach, one would normally build a model using time series methods such as trend, seasonal effects, autocorrelation […], and so on." (p. 7312)

Brodersen et al. (2015), Varian's colleagues at Google, developed a related approach. While not strictly necessary, they encourage analysts to include potential control groups (subjects that are similar to the treated subjects but unaffected by the intervention). In general, however, in their method, the only aspects that matter when choosing variables to estimate counterfactuals are pre-intervention correlation with the outcome variable of the treated subject and that they were not directly affected by the intervention. Example variables could be a similar but unaffected material, seasonality, or an index of general economic development.

6.3 The Action Design Research Methodology

We followed an action design research (ADR) process that was inspired by Sein et al. (Sein et al. 2011). ADR is a design research (DR) "method [that] reflects the premise that IT artifacts are ensembles shaped by the organizational context during development and use. The method conceptualizes the research process as containing the inseparable and inherently interwoven activities of building the IT artifact, intervening in the organization, and evaluating it concurrently." (p. 37) This is different from traditional DR approaches, such as the one proposed by Hevner et

al. (2004) that explicitly separate artifact design and evaluation. ADR consists of four main stages: "problem formulation," "building, intervention, and evaluation" (BIE), "reflection and learning," and "formalization of learning." (Sein et al. 2011)

Following the method, in the problem formulation stage, we identified and conceptualized the research opportunity, and formulated our research question for the class of value-based spare parts pricing support systems (see Section 6.1). Furthermore, we conceptualized and informed our initial solution design with a review of B2B marketing and causal inference literature (see Sections 6.1 and 6.2). Also, we set-up roles and responsibilities in the ADR team that, in its core, consisted of Oliver Müller as a researcher, Tiemo Thiess as both a researcher and developer and MAN Energy Solutions' pricing manager for the aftersales business as an end-user and key informant and evaluator of the system. We assured the long-term organizational commitment with a formal research collaboration agreement that was part of a larger ADR program that one of the researchers conducted at the company for his Ph.D. studies.

In the BIE stage, we built several prototypes before implementing the eventual system. Throughout this process, we continuously evaluated the artifacts. As our main criterion, we evaluated them in terms of their effectiveness, e.g., did the system help to solve the field problem? Also, we performed simple architectural analyses and white box tests (did the system execute without errors in the "technical infrastructure of the business environment?") (Hevner et al. 2004) Moreover, we performed functional black-box tests, e.g., did the model predict the pre-intervention data satisfyingly well? Moreover, we evaluated the artifacts based on end-user feedback from demonstrations during status updates and stakeholder presentations (see Table 1).

In the reflection, learning, and formalization stages, we formalized key learnings and design decisions as design principles (Section 6.5.1). To inform theorizing, design, evaluation, and understanding of the, we collected empirically rich observation notes of our main encounters at MAN Energy solutions (see Table 1). The data collection and analysis approach is an instance of design ethnography, a form of ethnography that relaxes strict objectivity assumptions and in-

stead focusses on studying the whole design situation in which artifacts, stakeholders, and the researchers itself dynamically interact with each other (Baskerville and Myers 2015). Also, we had access to many of the companies' systems, databases, and documents that further informed our study.

Table 1. Study-related encounters at MAN Energy Solutions

Meeting Type	Participants	#	h	Total (h)
Status Update	Pricing Manager, Researchers	7	1	7
Stakeholder Presentation	Diverse, including middle management	2	1	2
Development Sprint	Researchers	14	4	56

6.4 Design and Implementation of a Causal Inference System for Value-based Spare Parts Pricing at MAN Energy Solutions

After we conceptualized our initial solution class as causal inference systems, we developed throughout several iterations of building, intervention, and evaluation, different prototypes. The explicit goal was to design and implement a system that helps to evaluate the success of value-based pricing initiatives at MAN Energy Solutions. In particular, the system should allow to visualize the effects of value-based price changes on materials and contain additional material-specific data to aid learning, test hypotheses, and aid future pricing decisions. Moreover, the system should estimate the effects of value-based price changes for different key performance indicators, such as sales volume, sales revenue, and sales conversion rate.

Some of the early prototypes involved DID related methods; however, we quickly noticed their limitations when observing seemingly unrealistic estimates. Justified by the theoretical literature (Section 2), we quickly focused on prediction-based counterfactual estimation methods, as

described by Varian (2016). We tried and compared many approaches to estimate the counterfactual based on their ability to predict the pre-intervention data, and on plausibility checks. Such checks, again, were based on our and our informant's subject-matter knowledge. Among the approaches that we tried were generalized least squares (GLS), autoregressive integrated moving average (ARIMA), OLS, and Croston's method for intermittent demand (Croston 1972; Harvey and Amemiya 1987). However, eventually, an approach based on a Bayesian structural time-series (BSTS; Brodersen et al. 2015; Scott and Varian 2014) model consisting of a local-level trend and a regression component performed best.

To increase the predictability of the often irregular and "lumpy" aftersales data (Bartezzaghi et al. 1999), we tried a variety of different grouping characteristics. Grouping the treated materials based on their relative unit-price level (between 1 and 10), and the relative price change level (between -10 to +10) was the most plausible and best-performing approach. Also, we aggregated materials with unchanged prices into potential control groups based on their relative unit-price-level and a variable that indicates which regional headquarter is mainly responsible for selling it.

To predict counterfactual outcomes of the situation that one would have expected had the price not changed, we first had to process the raw transactional data and transform it into monthly time-series data.

In the BSTS model, we include a local-level trend component for the outcomes (e.g., average sales volume) of the treated unit (a repriced material group). Moreover, for the regression component, we include the following independent variables: 1.) a yearly seasonality term, 2.) a monthly seasonality term, 3.) the days of the month (e.g., 28 in February), 4.) Fourier terms (complex seasonality; see Hyndman and Athanasopoulos 2018), 5.) a variable of the average monthly unit price for the repriced material group that, for the post-intervention periods, we fixed at the last pre-intervention price, 6.) the potential material control groups.

As we fit the model on many variables, and in particular on many potential control groups, we include an automatic variable selection method. As described by Scott and Varian (2013) and

Brodersen et al. (2015), we place spike-and-slab priors on the regression coefficients to affect the probability that the algorithm includes a variable into the model. This way, the algorithm can exclude uncorrelated variables fully.

In simplified terms, the algorithm first fits a model or function $F(X) = Y$ on the pre-intervention data (when ignoring the local level trend component), then, the system inserts the observed values for the independent variables (e.g., X_1, X_2, X_3) in the post-intervention period to predict the counterfactual (Y_0). Then, the system subtracts the predicted counterfactual outcomes in the post-intervention period from the observed outcomes (Y_1) in the post-intervention period to calculate treatment effects (see Fig. 1 for an illustration).

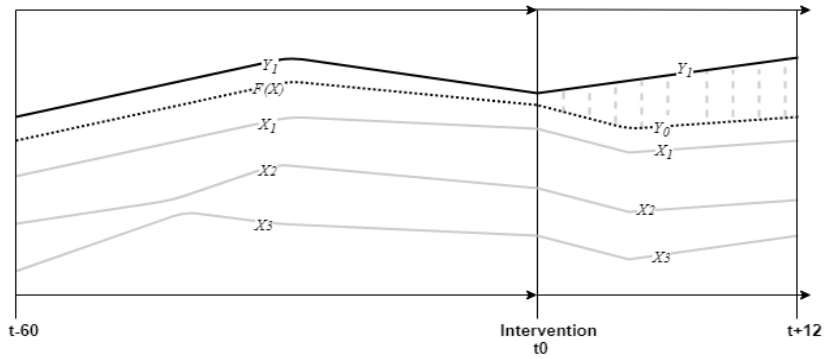

Figure 1. Counterfactual Prediction Approach

To assure that none of the control variables are affected by the intervention on the treated subject (a material group), we include with the engine family an "unaffectedness condition." In the system execution process (Figure 2), we first group the materials on the above-described characteristics and engine family. Once the system starts to loop through each material group, it can first filter all control variables of the same engine family out before further aggregating the materials into their eventual groups.

Also, before the system fits the model on all pre-intervention data, it performs time-series cross-validation to estimate average mean absolute prediction errors that indicate the pre-intervention predictability of a given repriced material group.

Eventually, the system represents the analysis outputs in an interactive report. The snapshot in Fig. 3 shows the sum of the actual outcomes (e.g., sales volume) for repriced material groups (red) that the user selects. It also shows the sum of the predicted counterfactual outcomes for such repriced material groups (graphite), and the sum of total treatment effects (e.g., sales volume; grey). The diagrams show the sums for the full post-intervention period (Y2019) and for each quarter separately. Via the sheets, one can choose the KPI of interest (e.g., sales volume, sales revenue, or conversion rate), and assess additional information about the materials and their respective groups.

The system was well-received by MAN Energy Solutions, and based on it, they could validate many of the hypotheses behind their recent value-based pricing decisions and actions. As a result of this, the top management decided to role the value-based pricing initiative, that at this point was only conducted for the product range of one regional headquarter, out to other regional headquarters and their product ranges as well.

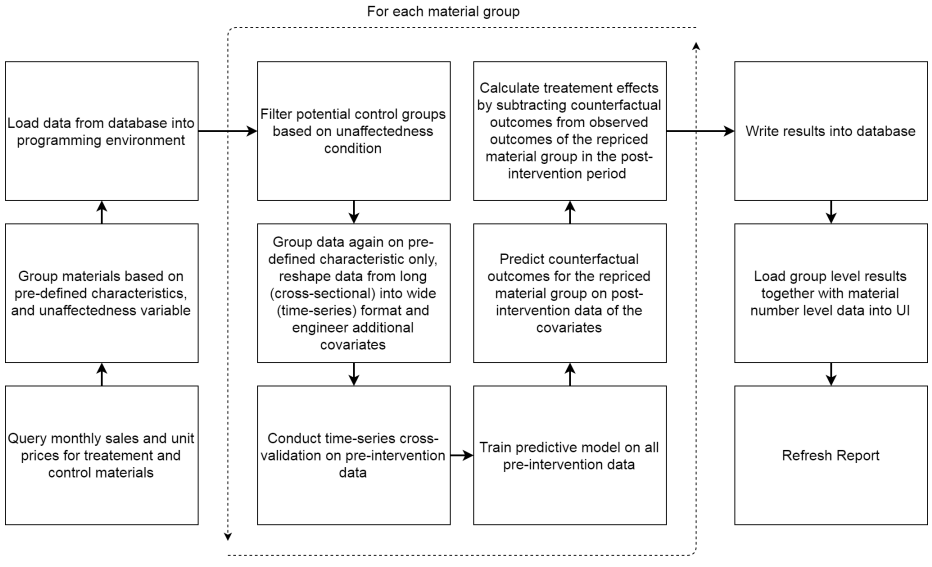

Figure 2. The system execution process

Figure 3. Interactive report with adjustable filters (blurred for confidentiality reasons)

6.5 Discussion of Design Principles

6.5.1 DP1: Pre-aggregation – analysts should pre-aggregate lumpy data to improve its predictability

In the spare parts business in general and at MAN in particular, one often has to deal with large portfolios of materials with lumpy demand patterns. The reasons for this are 1.) a relatively small number of customers for a particular material, 2.) a heterogeneous customer base, e.g., a few large and many small customers, 3.) infrequent purchases, e.g., due to engine life-cycle dependent spare parts, and 4.) variable requests, e.g., large orders in case of breakdowns or mayor overhauls followed by small orders without a continuous pattern (see Bartezzaghi et al. 1999; Bartezzaghi and Kalchschmidt 2011).

Research suggests data aggregation as a way to improve the validity and predictability of statistical models when dealing with lumpy demand data (Bartezzaghi et al. 1999; Bartezzaghi and Kalchschmidt 2011; Zotteri and Kalchschmidt 2007). According to Zotteri (2007), when one wants to implement predictive models in such a situation, "the problem of the aggregation level of data […] is much more complex than the simple design or selection of an appropriate algorithm, and it involves the choice of the relevant pieces of information, the design of information systems, the control of data quality, and the definition of managerial processes." (p. 7)

Similarly, we tried many different pre-aggregation characteristics and levels before we eventually settled on aggregating the data in terms of their temporal structure, e.g., aggregated daily to monthly data, and in terms of some pre-defined characteristics such as their relative unit-price-level and price-change-level as the best performing and most plausible approach.

6.5.2 DP2: Scalability – analysts should use robust algorithms that rely on few assumptions only and include global explainability features to enable controlled execution at scale

At MAN Energy Solutions, we had to estimate counterfactuals and calculate treatment effects for numerous material groups. Even though we reduced complexity by pre-aggregating the data into broader groups, we still had to conduct so many analyses that a thorough visual modeling approach was infeasible. Moreover, fitting models with many covariates is challenging because the amount of data needed to estimate model parameters accurately grows exponentially with each additional variable ("the curse of dimensionality"; Bellman 2015). This problem becomes even more severe when working with time-series data, as one often has only about a hundred records (monthly data) or less (lumpy data) per unit for model fitting.

We addressed the challenge by choosing a BSTS model. This modeling approach generally shows good performance on many different datasets and involves, with spike and slab priors, an automatic variable selection method (Scott and Varian 2013, 2014). Moreover, in our BSTS model, we chose a local-level trend component, which performs well in many cases (Brodersen et al. 2015). Furthermore, we did not include seasonality as a state component of the BSTS model. Still, we included seasonality terms as covariates into the model, so that the variable selection method can include them whenever they help to explain variance in the outcome variable of a repriced material group and exclude them when they don't. This approach is more flexible and requires fewer assumptions about the unobserved state-space (data generating process) than pre-selecting a seasonality state component for all repriced material groups.

Also, we struggled to understand the behavior of individual models fully, as a visual exploration of all fitted models for all repriced material groups was simply not possible. Because of that, we implemented explainability features to help us understand better and improve

our models. In particular, we displayed the strongest predictor together with its inclusion probability for each model and calculated pre-intervention prediction errors, respectively. Research about global explainability methods justifies our approach (Gregor and Benbasat 1999; Martens and Provost 2014; Thiess et al. 2020).

6.5.3 DP3: Unaffectedness – analysts should define unaffectedness conditions based on subject-matter knowledge, and causal diagrams and filter model covariates based on them to avoid spillover effects

The most important assumption in synthetic control based causal inference is that the control groups are not directly affected by the treatment (Abadie et al. 2010; Abadie and Gardeazabal 2003; Brodersen et al. 2015). To assure that one does not violate this assumption, one could simply choose all non-treated material groups as potential controls. In aftersales and at MAN Energy Solutions, however, this approach can lead to biased estimates as spare parts orders can be large and consist of several combinations of frequently bought together materials. The result of this is that a price change in one of the related materials can affect the demand for the other material, which is equivalent to a spillover effect.

We addressed this challenge by systematically mapping all potential causes of spillover effects using our and our informant's subject-matter knowledge as well as causal diagram-like representations. We justify our approach by referring to Pearl (1995) and Hernan et al. (2002) that suggest both the use of subject-matter knowledge and causal diagrams for causal reasoning and modeling.

Based on such an approach, we defined the engine family of a material as an appropriate unaffectedness condition and filtered potential control groups based on it. We made this decision based on the insight that it is unlikely that customers frequently buy materials of different engine families together because they usually order spare parts for a particular ship that usually only has one or two main engines of the same family installed.

6.5.4 DP4: Pre-intervention predictability – analysts should use cross-validation and evaluate treatment effects in light of the pre-intervention predictability to draw more truthful conclusions

According to Brodersen et al. (2015), the main assumptions that causal impact analyses need to fulfill are 1.) the unaffectedness of controls by the treatment and 2.) a relatively strong pre-intervention correlation of the covariates with the pre-intervention outcome of the treated unit. The first assumption is essential to avoid biased estimates due to spillover effects and the second one is important to assure a good model fit, and, thus, a valid estimation of the counterfactual. Many researchers have convincingly shown that goodness of fit is not sufficient in analyses that require predictive estimates (Fildes and Makridakis 1995; Makridakis et al. 1982; Tashman 2000). Moreover, some time-series are simply easier to predict than others, such as time-series that have regular seasonality patterns and stable trends. At MAN Energy Solutions, we had to work with lumpy time-series that, despite the increased data quality due to pre-aggregation, still did not fulfill such ideal properties.

In reaction to this, we assessed the pre-intervention predictability in terms of mean absolute percentage error (MAPE) for outcome data of each repriced material group by performing time-series cross-validation (Bergmeir et al. 2018). We then included a condition that marks the mean absolute percentage treatment effect for each treated material group as significant only if they are larger than the pre-intervention prediction MAPE. By doing so, we avoid displaying treatment effects that are only due to the variation in pre-intervention predictability.

6.5.5 DP5: Treatment simulation – analysts should add the treatment variable to the model and fix its post-intervention values at its last pre-intervention value to strengthen the counterfactual prediction

At the root of causal inference is the potential outcome framework by Donal Rubin (Rubin 2005) that suggests estimating causal treatment effects as the difference between the observed and the (counterfactual) potential outcome. All of the causal inference methods discussed above apply counterfactual reasoning as they try to simulate a world in which the unobservable potential outcome was observed. So they try to answer the question: "How would the outcome be had the treatment not occurred?" (Morgan and Winship 2015; Pearl 1999)

In our case, the treatment of interest was a price change at a particular point in time. Nevertheless, it was likely not the first price change for a given material. In our model, we explain histor-

ical variations in the outcome that were caused by prior treatments by incorporating a continuous treatment variable (the monthly average unit price) as a predictor directly into the model. For the post-intervention prediction, however, we fix its values at the last pre-intervention unit price. By doing so, we account for the effects of prior interventions on the outcome time-series. Moreover, as we insert simulated (fixed) values for the treatment variable, we strengthen the counterfactual prediction of the model by using it as a simulation engine directly.

6.5.6 DP6: Interactive visualization – analysts should create interactive reports instead of static presentations to aid understanding and acceptance

During our ADR project, we reasoned that it could be difficult for users to comprehend the inner workings of our analysis approach. Also, the pricing manager that was the key end-user of the system requested a possibility to explore the data interactively. Research in information visualization suggests that interactive representations of data increase the usability of systems and aid learning and understanding for its users (Liu et al. 2014). Research in technology acceptance, on the other hand, suggests that an increased understanding increases the acceptance of a decision support system (Gregor and Benbasat 1999; Kayande et al. 2009; Martens and Provost 2014).

Motivated by these findings, we designed an interactive user interface in which users can explore the causal impact of price changes on different KPIs, e.g., by adjusting the information that the diagrams represent via slicers for the unit-price-level or the price-change-level of materials. Here, whenever one adjusts a slicer, the visualization adapts immediately.

6.6 Discussion and Conclusions

In this study, we contribute to information systems and industrial marketing by answering our research question with theory about how to design and implement causal inference systems for value-based spare parts pricing support (Gregor and Jones 2007). Our study provides theoretical and empirical evidence on how causal inference approaches can solve the relevant field problem of estimating the causal effects of pricing interventions not only on subjects in a laboratory setting but on the experimenter (the firm) in its natural context (Varian 2016). Moreover, we show in a particular industry application how artificial intelligence (AI) related technologies

can foster servitization processes towards more value-based and customer-centric sales and marketing approaches (Huang and Rust 2018).

While our system incorporates an approach for counterfactual prediction and treatment effect calculation that is comparable to the one introduced by Brodersen et al. (2015), our approach is specially adapted to the aftersales context and value-based pricing effect estimation problems (see DP1, DP2, and DP3). Moreover, our approach is an industrial application that required us to design and implement a socio-technical information system around the causal inference approach that is fully integrated into the IT architecture at MAN Energy Solution and includes a data processing pipeline and a user interface (see DP1 and DP6). Brodersen et al., on the other hand, introduce a general-purpose method for causal impact analysis and demonstrate its utility in a laboratory setting on empirical data from a digital marketing campaign.

Furthermore, we improve the method by Brodersen et al. by adding measures of *pre-intervention predictability* (DP4) and adding a treatment variable to the model (unit price) that helps to explain the variation in the pre-intervention period and strengthens the counterfactual prediction by fixing it at its last pre-intervention value for the post-intervention period (DP5).

In the future, we want to further experiment with pre-aggregation levels to find a way that balances model robustness with more fine-grained applicability of results in organizational decision making processes. Furthermore, we want to investigate how time-series cross-validation based hyperparameter tuning and prediction combinations affect system performance and treatment-effect-validity.

References

van der Aalst, W. M. P. 2014. “Data Scientist: The Engineer of the Future,” in *Enterprise Interoperability VI*, K. Mertins, F. Bénaben, R. Poler, and J.-P. Bourrières (eds.), Springer International Publishing, pp. 13–26. (https://doi.org/10.1007/978-3-319-04948-9_2).

Abadie, A. 2005. “Semiparametric Difference-in-Differences Estimators,” *The Review of Economic Studies* (72:1), Wiley-Blackwell, pp. 1–19. (https://doi.org/10.1111/0034-6527.00321).

Abadie, A., Diamond, A., and Hainmueller, J. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program,” *Journal of the American Statistical Association* (105:490), Taylor & Francis, pp. 493–505. (https://doi.org/10.1198/jasa.2009.ap08746).

Abadie, A., and Gardeazabal, J. 2003. “The Economic Costs of Conflict: A Case Study of the Basque Country,” *American Economic Review* (93:1), pp. 113–132.

Abadie, A., and Imbens, G. W. 2006. “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica* (74:1), Wiley Online Library, pp. 235–267. (https://doi.org/10.1111/j.1468-0262.2006.00655.x).

Abe, M. 2008. “‘Counting Your Customers’ One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model,” *Marketing Science* (28:3), pp. 541–553. (https://doi.org/10.1287/mksc.1090.0502).

Adrodegari, F., Alghisi, A., and Saccani, N. 2014. “Towards Usage-Oriented Business Models : An Assessment of European Capital Goods Manufacturers Abstract :,” *21st Euroma Conference* (July 2016), pp. 1–10.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., and Rush, J. D. 2006. “The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction,” *The Counseling Psychologist* (34:3), Sage PublicationsSage CA: Thousand Oaks, CA, pp. 341–382. (https://doi.org/10.1177/0011000005285875).

Ågerfalk, P. J. 2020. “Artificial Intelligence as Digital Agency,” *European Journal of Information Systems* (29:1), pp. 1–8.

Aken, J. E. van. 2004. “Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules,” *Journal of Management Studies* (41:2), Wiley Online Library, pp. 219–246.

Andersson, J., and Bengtsson, J. 2013. “Spare Parts Pricing-Pre-Study for a Pricing Strategy at Pon.”

Arnott, D., and Pervan, G. 2008. “Eight Key Issues for the Decision Support Systems Discipline,” *Decision Support Systems* (44:3), pp. 657–672. (https://doi.org/10.1016/j.dss.2007.09.003).

Arnott, D., and Pervan, G. 2014. “A Critical Analysis of Decision Support Systems Research Revisited: The Rise of Design Science,” *Journal of Information Technology* (29:4), pp. 269–293.

(https://doi.org/10.1057/jit.2014.16).

Ascarza, E., Fader, P. S., and Hardie, B. G. S. 2017. "Marketing Models for the Customer-Centric Firm," in *International Series in Operations Research and Management Science* (Vol. 254), Springer, pp. 297–329. (https://doi.org/10.1007/978-3-319-56941-3_10).

Ashenfelter, O., and Card, D. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics* (Vol. 67). (https://doi.org/10.2307/1924810).

Avison, D. E., and Fitzgerald, G. 1995. *Information Systems Development: Methodologies, Techniques and Tools, Paul & Company Publishers Consortium*, Inc.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. 2010. "How to Explain Individual Classification Decisions," *The Journal of Machine Learning Research* (11), JMLR. org, pp. 1803–1831.

Baines, T., Bigdeli, A. Z., Bustinza, O. F., Shi, V. G., Baldwin, J., and Ridgway, K. 2017. "Servitization: Revisiting the State-of-the-Art and Research Priorities," *International Journal of Operations & Production Management*, Emerald Publishing Limited.

Baines, T., and Lightfoot, H. 2012. "Made to Serve," *Made to Serve*, Wiley Online Library. (https://doi.org/10.1002/9781119207955).

Bartezzaghi, E., and Kalchschmidt, M. 2011. "The Impact of Aggregation Level on Lumpy Demand Management," in *Service Parts Management*, Springer, pp. 89–104.

Bartezzaghi, E., Verganti, R., and Zotteri, G. 1999. "Simulation Framework for Forecasting Uncertain Lumpy Demand," *International Journal of Production Economics* (59:1), Elsevier Science B.V., pp. 499–510. (https://doi.org/10.1016/S0925-5273(98)00012-7).

Baskerville, R. L., and Myers, M. D. 2015. "Design Ethnography in Information Systems," *Information Systems Journal* (25:1), John Wiley & Sons, Ltd (10.1111), pp. 23–46.

Baskerville, R., and Wood-Harper, T. 1996. "A Critical Perspective on Action Research as a Method for Information Systems Knowledge Transfer Practice View Project Development of Multiview Using Action Research View Project," *Article in Journal of Information Technology*. (https://doi.org/10.1080/026839696345289).

Bazerman, M. H., and Moore, D. A. 2008. *Judgment in Managerial Decision Making*, Wiley.

Bellman, R. 1978. *An Introduction to Artificial Intelligence: Can Computers Think?*, Thomson Course Technology.

Bellman, R. E. 2015. *Adaptive Control Processes: A Guided Tour*, Princeton university press.

Benbasat, I., and Zmud, R. W. 1999. "Empirical Research in Information Systems: The Practice of Relevance," *MIS Quarterly*, JSTOR, pp. 3–16.

Benbasat, I., and Zmud, R. W. 2003. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties," *MIS Quarterly: Management Information*

Systems (27:2), JSTOR, pp. 183–194. (https://doi.org/10.2307/30036527).

Benbasat, I., and Zmud, R. W. 2006. "Empirical Research in Information Systems: The Practice of Relevance," *MIS Quarterly* (23:1), p. 3.

Berente, N., Seidel, S., and Safadi, H. 2019. "Data-Driven Computationally-Intensive Theory Development."

Bergmeir, C., Hyndman, R. J., and Koo, B. 2018. "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction," *Computational Statistics & Data Analysis* (120), Elsevier, pp. 70–83.

Berry, D. C., and Broadbent, D. E. 1987. "Explanation and Verbalization in a Computer-Assisted Search Task," *The Quarterly Journal of Experimental Psychology Section A* (39:4), Taylor & Francis, pp. 585–609. (https://doi.org/10.1080/14640748708401804).

Bertrand, M., Duflo, E., and Mullainathan, S. 2004. "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics* (119:1), MIT Press, pp. 249–275.

Beverungen, D., Matzner, M., and Janiesch, C. 2017. *Information Systems for Smart Services*, Springer.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, Oxford university press.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*, springer.

Blattberg, R. C., Getz, G., and Thomas, J. S. 2001. *Customer Equity: Building and Managing Relationships as Valuable Assets (Harvard Business School Press, Boston, Massachusetts)*.

Bohanec, M., Kljajić Borštnar, M., and Robnik-Šikonja, M. 2017. "Explaining Machine Learning Models in Sales Predictions," *Expert Systems with Applications* (71), Pergamon, pp. 416–428.

Bohanec, M., Robnik-Šikonja, M., and Kljajić Borštnar, M. 2017a. "Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting," *Organizacija* (50:3), pp. 217–233.

Bohanec, M., Robnik-Šikonja, M., and Kljajić Borštnar, M. 2017b. "Decision-Making Framework with Double-Loop Learning through Interpretable Black-Box Machine Learning Models," *Industrial Management & Data Systems* (117:7), Emerald Publishing Limited, pp. 1389–1406.

Boyd, D., and Crawford, K. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* (15:5), Taylor & Francis, pp. 662–679.

Božič, K., and Dimovski, V. 2019. "Business Intelligence and Analytics for Value Creation: The Role of Absorptive Capacity," *International Journal of Information Management* (46), Elsevier, pp. 93–103.

Breiman, L. 1997. "Arcing the Edge," *Statistics* (4), pp. 1–14.

Breiman, L. 2001a. "Random Forests," *Machine Learning* (45:1), pp. 5–32. (https://doi.org/10.1023/A:1010933404324).

Breiman, L. 2001b. "Statistical Modeling: The Two Cultures," *Statistical Science* (Vol. 16).

vom Brocke, J., Debortoli, S., Müller, O., and Reuter, N. 2014. "How In-Memory Technology Can Create Business Value: Insights from the Hilti Case," *Communications of the Association for Information Systems*.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. 2015. "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *Annals of Applied Statistics* (9:1), pp. 247–274. (https://doi.org/10.1214/14-AOAS788).

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. 2011. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?," *SSRN Electronic Journal*. (https://doi.org/10.2139/ssrn.1819486).

Brynjolfsson, E., and Mcafee, A. 2017. "The Business of Artificial Intelligence: What It Can--and Cannot--Do for Your Organization," *Harvard Business Review Digital Articles* (7), pp. 3–11.

Brynjolfsson, E., and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York City, NY, USA: WW Norton & Company.

Brynjolfsson, E., and Mitchell, T. 2017. "What Can Machine Learning Do? Workforce Implications: Profound Change Is Coming, but Roles for Humans Remain," *Science* (358:6370), American Association for the Advancement of Science, pp. 1530–1534. (https://doi.org/10.1126/science.aap8062).

Brynjolfsson, E., Rock, D., and Syverson, C. 2017. "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," *NBER Working Paper 24001*.

Card, D. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* (43:2), SAGE Publications Sage CA: Los Angeles, CA, p. 245. (https://doi.org/10.2307/2523702).

Cartwright, N. 2007. "Are RCTs the Gold Standard?," *BioSocieties* (2:1), Cambridge University Press, pp. 11–20. (https://doi.org/10.1017/s1745855207005029).

Cartwright, N. 2009. "Evidence-Based Policy: What's to Be Done about Relevance?," *Philosophical Studies* (143:1), Springer, pp. 127–136.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, JSTOR, pp. 1165–1188.

Cochrane, A. L. 1972. *Effectiveness and Efficiency: Random Reflections on Health Services*, (Vol. 900574178), Nuffield Provincial Hospitals Trust London.

Cohen, M. A., Agrawal, N., and Agrawal, V. 2006. "Winning in the Aftermarket," *Harvard Business Review* (84:5), pp. 129–138.

Cook, T. D., Campbell, D. T., and Day, A. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, (Vol. 351), Houghton Mifflin Boston.

Coombs, C., Hislop, D., Taneva, S. K., and Barnard, S. 2020. "The Strategic Impacts of Intelligent Automation for Knowledge and Service Work: An Interdisciplinary Review," *The Journal of Strategic Information Systems*, p. 101600. (https://doi.org/10.1016/j.jsis.2020.101600).

Cowan, N. 2001. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity,” *Behavioral and Brain Sciences* (24:1), pp. 87–114.

Craven, M., and Shavlik, J. W. 1996. “Extracting Tree-Structured Representations of Trained Networks,” in *Advances in Neural Information Processing Systems*, pp. 24–30.

Croston, J. D. 1972. “Forecasting and Stock Control for Intermittent Demands.,” *Operational Research Quarterly* (23:3), Springer, pp. 289–303. (https://doi.org/10.1057/jors.1972.50).

Cullbrand, M., and Levén, L. 2012. *Spare Parts Pricing-Setting the Right Prices for Sustainable Profit at Atlet*, Chalmers University of Technology.

D’Haen, J., and Van den Poel, D. 2013. “Model-Supported Business-to-Business Prospect Prediction Based on an Iterative Customer Acquisition Framework,” *Industrial Marketing Management* (42:4), Elsevier, pp. 544–551.

Daniel, E. M., Ward, J. M., and Franken, A. 2014. “A Dynamic Capabilities Perspective of IS Project Portfolio Management,” *The Journal of Strategic Information Systems* (23:2), pp. 95–111. (https://doi.org/10.1016/j.jsis.2014.03.001).

Danish Ship Finance. 2018. “Shipping Market Review.”

Daugherty, P. R., and Wilson, H. J. 2018. *Human+ Machine: Reimagining Work in the Age of AI*, Harvard Business Press.

Davenport, T. H. 2013. “Analytics 3.0,” *Harvard Business Review* (91:12), HARVARD BUSINESS SCHOOL PUBLISHING CORPORATION 300 NORTH BEACON STREET, WATERTOWN, MA 02472 USA, pp. 64--+.

Davenport, T. H., and Harris, J. G. 2007. *Competing on Analytics: The New Science of Winning*, Harvard Business Press.

Davenport, T. H., and Kirby, J. 2016. *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*, Harper Business New York, NY.

Davenport, T. H., and Patil, D. J. 2012. “Data Scientist: The Sexiest Job of the 21st Century,” *Harvard Business Review*.

Davenport, T. H., and Ronanki, R. 2018. “Artificial Intelligence for the Real World,” *Harvard Business Review* (January-Fe:1), pp. 108–116.

Davis, G. B. 2005. “Advising and Supervising,” *Research in Information Systems - A Handbook for Research Supervisors and Their Students*, Amsterdam: Elsevier Butterworth Heinemann, pp. 3–34.

Davison, R. M., Martinsons, M. G., and Kock, N. 2004. “Principles of Canonical Action Research,” *Information Systems Journal* (14:1), Wiley Online Library, pp. 65–86. (https://doi.org/10.1111/j.1365-2575.2004.00162.x).

Dawes, R. M. 1979. “The Robust Beauty of Improper Linear Models in Decision Making.,” *American Psychologist* (34:7), American Psychological Association, p. 571.

Dawes, R. M., Faust, D., and Meehl, P. E. 1989. “Clinical versus Actuarial Judgment,” *Science* (243:4899), American Association for the Advancement of Science, pp. 1668–1674.

Dearden, A. 2001. “IDA-S: A Conceptual Framework for Partial Automation,” in *People and Computers XV—Interaction without Frontiers*, Springer, London, pp. 213–228.

Dhaliwal, J. S. 1993. “An Experimental Investigation of the Use of Explanations Provided by Knowledge-Based Systems,” University of British Columbia. (https://doi.org/10.14288/1.0086429).

De Dreu, C. K. W., Beersma, B., Stroebe, K., and Euwema, M. C. 2006. “Motivated Information Processing, Strategic Choice, and the Quality of Negotiated Agreement,” *Journal of Personality and Social Psychology* (90:6), American Psychological Association, pp. 927–943. (https://doi.org/10.1037/0022-3514.90.6.927).

Drucker, P. F. 1967. “The Effective Decision,” *Harvard Business Review*, pp. 28–34.

Druzdzel, M. J., and Díez, F. J. 2003. “Combining Knowledge from Different Sources in Causal Probabilistic Models,” *Journal of Machine Learning Research* (4:Jul), pp. 295–316.

Du, M., Liu, N., and Hu, X. 2019. “Techniques for Interpretable Machine Learning,” *Communications of the ACM* (63:1), ACM New York, NY, USA, pp. 68–77.

Du, M., Liu, N., and Hu, X. 2020. “Techniques for Interpretable Machine Learning,” *Communications of the ACM* (63:1), ACM New York, NY, USA, pp. 68–77. (https://doi.org/10.1145/3359786).

Duncan, B., and Elkan, C. 2015. “Probabilistic Modeling of a Sales Funnel to Prioritize Leads,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015-Augus), pp. 1751–1758.

Dybowski, R., Laskey, K. B., Myers, J. W., and Parsons, S. 2003. “Introduction to the Special Issue on the Fusion of Domain Knowledge with Data for Decision Support,” *The Journal of Machine Learning Research* (4), JMLR. org, pp. 293–294.

Edwards, W. 1954. “The Theory of Decision Making,” *Psychological Bulletin* (51:4), American Psychological Association, pp. 380–417. (https://doi.org/10.1037/h0053870).

Efron, B., and Morris, C. 1977. “Stein’s Paradox in Statistics,” *Scientific American* (236:5), pp. 119–127.

Einhorn, H. J., and Hogarth, R. M. 1985. “Ambiguity and Uncertainty in Probabilistic Inference,” *Psychological Review* (92:4), American Psychological Association, pp. 433–461. (https://doi.org/10.1037/0033-295X.92.4.433).

Eisenhardt, K. M. 1989. “Building Theories from Case Study Research,” *The Academy of Management Review* (14:4), pp. 532–550. (https://doi.org/10.2307/258557).

Eitle, V., and Buxmann, P. 2019. “Business Analytics for Sales Pipeline Management in the Software Industry: A Machine Learning Perspective,” *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

European Union. 2018. *Recital 71 - Profiling | General Data Protection Regulation (GDPR)*.

Evans, J. S. B. T. 2006. “The Heuristic-Analytic Theory of Reasoning: Extension and Evaluation,” *Psychonomic Bulletin & Review* (13:3), Springer, pp. 378–395.

Everett, A. M. 1995. *An Empirical Investigation of the Effect of Variations in Expert System Explanation*

Presentation on Users' Acquisition of Expertise and Perceptions of the System.

Fader, P. 2012. *Customer Centricity : Focus on the Right Customers for Strategic Advantage*, Wharton digital press.

Fader, P. S., and Hardie, B. G. 2013. *The Gamma-Gamma Model of Monetary Value*.

Fader, P. S., and Hardie, B. G. S. 2009. "Probability Models for Customer-Base Analysis," *Journal of Interactive Marketing* (23:1), Anniversary Issue, pp. 61–69. (https://doi.org/10.1016/j.intmar.2008.11.003).

Fader, P. S., Hardie, B. G. S., and Lee, K. L. 2005. "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science* (24:2), INFORMS, pp. 275–284.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62–70. (https://doi.org/10.1016/j.infoandorg.2018.02.005).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* (17:3), pp. 37–53.

Festinger, L. 1957. *A Theory of Cognitive Dissonance*, (Vol. 2), Stanford university press.

Fildes, R., and Makridakis, S. 1995. "The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting," *International Statistical Review/Revue Internationale de Statistique*, JSTOR, pp. 289–308.

Floridi, L., and Sanders, J. W. 2004. "On the Morality of Artificial Agents," *Minds and Machines* (14:3), pp. 349–379.

Fountaine, T., McCarthy, B., and Saleh, T. 2019. "Building the AI-Powered Organization," *Harvard Business Review* (97:4), pp. 62–73.

Van Fraassen, B. C. 1980. *The Scientific Image*, Oxford University Press.

François Chollet. 2017. "Deep Learning with Python," *Deep Learning with Python*, p. 386.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. 2019. "Toward Understanding the Impact of Artificial Intelligence on Labor," *Proceedings of the National Academy of Sciences* (116:14), pp. 6531–6539. (https://doi.org/10.1073/pnas.1900949116).

Freund, Y., and Schapire, R. E. 1996. "Experiments with a New Boosting Algorithm."

Friedman, B. J. H. 2001. "1999 Reitz Lecture - Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics* (29:5), pp. 1189–1232.

Friedman, J. H. 2002. "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis* (38:4), North-Holland, pp. 367–378. (https://doi.org/10.1016/S0167-9473(01)00065-2).

Friedman, J., Hastie, T., and Tibshirani, R. 2001. *The Elements of Statistical Learning*, (Vol. 1), Springer series in statistics New York.

Gallagher, T., Mitchke, M. D., and Rogers, M. C. 2005. "Profiting from Spare Parts," *The McKinsey Quarterly* (2:Exhibit 2), pp. 1–4.

Galliers, R. D., Newell, S., Shanks, G., and Topi, H. 2017. "Datification and Its Human, Organizational and Societal Effects: The Strategic Opportunities and Challenges of Algorithmic Decision-Making," *The Journal of Strategic Information Systems* (26:3), pp. 185–190. (https://doi.org/10.1016/j.jsis.2017.08.002).

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. 2014. "A Survey on Concept Drift Adaptation," *ACM Computing Surveys* (46:4), ACM New York, NY, USA, pp. 1–37. (https://doi.org/10.1145/2523813).

Garfinkel, A. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory*, New Haven: Yale University Press.

Gigerenzer, G. 2008. "Why Heuristics Work," *Perspectives on Psychological Science* (3:1), SAGE Publications Sage CA: Los Angeles, CA, pp. 20–29.

Gioia, D. A., Corley, K. G., and Hamilton, A. L. 2013. "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology," *Organizational Research Methods* (16:1), Sage Publications Sage CA: Los Angeles, CA, pp. 15–31. (https://doi.org/10.1177/1094428112452151).

Glaser, B. G., and Strauss, A. L. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, (4. paperba.), Chicago: Aldine Publishing.

Goldkuhl, G. 2002. "Anchoring Scientific Abstractions – Ontological and Linguistic Determination Following Socio-Instrumental Pragmatism," in *European Conference on Research Methods in Business and Management*, pp. 1–30.

Goodacre, S. 2015. "Uncontrolled Before-after Studies: Discouraged by Cochrane and the EMJ," *Emergency Medicine Journal* (32:7), BMJ Publishing Group Ltd and the British Association for Accident~…, pp. 507–508. (https://doi.org/10.1136/emermed-2015-204761).

Goodwin, P., and Wright, G. 2014. *Decision Analysis for Management Judgment 5th Ed*, John Wiley & Sons Inc.

Greenland, S., Pearl, J., and Robins, J. M. 1999. "Causal Diagrams for Epidemiologic Research," *Epidemiology* (10:1), JSTOR, pp. 37–48. (https://doi.org/10.1097/00001648-199901000-00008).

Gregor, S. 2002. "Design Theory in Information Systems," *Australasian Journal of Information Systems* (10:1). (https://doi.org/10.3127/ajis.v10i1.439).

Gregor, S. 2006. "The Nature of Theory in Information Systems," *MIS Quarterly*, JSTOR, pp. 611–642.

Gregor, S., and Benbasat, I. 1999. "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," *MIS Quarterly: Management Information Systems* (23:4), Management Information Systems Research Center, University of Minnesota, pp. 497–530. (https://doi.org/10.2307/249487).

Gregor, S., Chandra Kruse, L., Seidel, S., and Kruse, C. 2020. "The Anatomy of a Design Principle The Anatomy of a Design Principle The Anatomy of a Design Principle* The Anatomy of a Design

Principle," *Article in Journal of the Association for Information Systems*.

Gregor, S. D. 1996. *Explanations from Knowledge-Based Systems for Human Learning and Problem Solving*.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly: Management Information Systems* (Vol. 37), University of Minnesota. (https://doi.org/10.25300/MISQ/2013/37.2.01).

Gregor, S., and Iivari, J. 2007. "Designing for Mutability in Information Systems Artifacts," *Information Systems Foundations: Theory, Representation and Reality*, pp. 3–24. (https://doi.org/10.22459/isftrr.11.2007.01).

Gregor, S., and Jones, D. 2007. "The Anatomy of a Design Theory," *Journal of the Association of Information Systems* (8:5), pp. 312–335.

Grønsund, T., and Aanestad, M. 2020. "Augmenting the Algorithm: Emerging Human-in-the-Loop Work Configurations," *The Journal of Strategic Information Systems*, p. 101614. (https://doi.org/10.1016/j.jsis.2020.101614).

Grove, W. M., and Meehl, P. E. 1996. "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical--Statistical Controversy.," *Psychology, Public Policy, and Law* (2:2), American Psychological Association, p. 293.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. 2000. "Clinical versus Mechanical Prediction: A Meta-Analysis.," *Psychological Assessment* (12:1), American Psychological Association, p. 19.

Grover, V., Chiang, R. H. L., Liang, T.-P., and Zhang, D. 2018. "Creating Strategic Business Value from Big Data Analytics: A Research Framework," *Journal of Management Information Systems* (35:2), pp. 388–423. (https://doi.org/10.1080/07421222.2018.1451951).

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., Feldberg, F., Mehrizi, M. H. R., Huysman, M., and Feldberg, F. 2017. "Debating Big Data: A Literature Review on Realizing Value from Big Data," *The Journal of Strategic Information Systems* (26:3), Elsevier, pp. 191–209. (https://doi.org/10.1016/J.JSIS.2017.07.003).

Gupta, S., and Lehmann, D. R. 2005. *Managing Customers as Investments. Wharton School Publishing*, FT Press Upper Saddle River.

Haj-Bolouri, A., Purao, S., Rossi, M., and Bernhardsson, L. 2017. "Action Design Research as a Method-in-Use: Problems and Opportunities," *KIT Scientific Working Papers* (Vol. 64).

Harman, G. H. 1965. "The Inference to the Best Explanation," *The Philosophical Review* (74:1), JSTOR, p. 88.

Hartl, E., and Hess, T. 2019. "It Projects in Digital Transformation: A Socio-Technical Journey Towards Technochange," in *ECIS 2019 Proceedings*, AIS electronic library.

Harvey, A. C., and Amemiya, T. 1987. "Advanced Econometrics.," *Economica* (Vol. 54), Harvard university press. (https://doi.org/10.2307/2554459).

Haussler, A. B. A. E. D., and Warmuth, M. K. 1987. “Occam’s Razor,” *Information Processing Letters* (24), pp. 377–380.

Hernán, M. A., and Cole, S. R. 2009. “Invited Commentary: Causal Diagrams and Measurement Bias,” *American Journal of Epidemiology* (170:8), Oxford University Press, pp. 959–962. (https://doi.org/10.1093/aje/kwp293).

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. 2004. “A Structural Approach to Selection Bias,” *Epidemiology* (15:5), JSTOR, pp. 615–625. (https://doi.org/10.1097/01.ede.0000135174.63482.43).

Hernán, M. A., Hernández-Diaz, S., Werler, M. M., and Mitchell, A. A. 2002. “Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology,” *American Journal of Epidemiology* (155:2), Oxford University Press, pp. 176–184. (https://doi.org/10.1093/aje/155.2.176).

Hernan, M. A., and Robins, J. M. 2010. *Causal Inference*, CRC Boca Raton, FL;

Hevner, A. 2007. “A Three Cycle View of Design Science Research,” *Scandinavian Journal of Information Systems* (19:2).

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. “Design Science in Information Systems Research,” *MIS Quarterly* (28:1), Management Information Systems Research Center, University of Minnesota, pp. 75–105.

Hinterhuber, A. 2004. “Towards Value-Based Pricing—An Integrative Framework for Decision Making,” *Industrial Marketing Management* (33:8), Elsevier, pp. 765–778.

Hinterhuber, A. 2008. “Value Delivery and Value-Based Pricing in Industrial Markets,” *Advances in Business Marketing and Purchasing* (14:1), pp. 381–448. (https://doi.org/10.1016/S1069-0964(08)14011-X).

Hinterhuber, A., and Liozu, S. M. 2014. “Is Innovation in Pricing Your next Source of Competitive Advantage?,” *Business Horizons* (57:3), Elsevier Ltd, pp. 413–423. (https://doi.org/10.1016/j.bushor.2014.01.002).

Hollander, E. P., Vroom, V. H., and Yetton, P. W. 1973. “Leadership and Decision-Making.,” *Administrative Science Quarterly*.

Huang, M.-H., and Rust, R. T. 2017. “Technology-Driven Service Strategy,” *Journal of the Academy of Marketing Science* (45:6), Springer, pp. 906–924.

Huang, M. H., and Rust, R. T. 2018. “Artificial Intelligence in Service,” *Journal of Service Research* (21:2), SAGE Publications Sage CA: Los Angeles, CA, pp. 155–172. (https://doi.org/10.1177/1094670517752459).

Hyndman, R. J., and Athanasopoulos, G. 2018. *Forecasting: Principles and Practice*, OTexts.

Iivari, J. 2003. “The IS Core-VII: Towards Information Systems as a Science of Meta-Artifacts,” *Communications of the Association for Information Systems* (12:1), p. 37.

Iivari, J. 2007. “A Paradigmatic Analysis of Information Systems As a Design Science,” *Scand. J. Inf. Syst.*

(19), p. 5.

International Data Corporation. 2019. “Worldwide Spending on 3D Printing Will Reach $13.8 Billion in 2019, According to New IDC Spending Guide,” *Press Release, IDC*.

International Maritime Organization. 2015. “SURVEY GUIDELINES UNDER THE HARMONIZED SYSTEM OF SURVEY AND CERTIFICATION (HSSC).”

Izenman, A. J. 1991. “Recent Developments in Nonparametric Density Estimation,” *Journal of the American Statistical Association* (86:413), Taylor & Francis, Ltd.American Statistical Association, p. 205.

Janik, A., Rieke, R., and Toulmin, S. 1984. *An Introduction to Reasoning*, New York, New York: MacMillan Publishing Co.

Janis, I. L., and Mann, L. 1977. *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment.*, Free press.

Jarrahi, M. H. 2019. “In the Age of the Smart Artificial Intelligence: AI’s Dual Capacities for Automating and Informating Work,” *Business Information Review* (36:4), pp. 178–187.

Josephson, J. R., and Josephson, S. G. 1996. *Abductive Inference: Computation, Philosophy, Technology*, Cambridge University Press.

Kahneman, D. 2003. “A Perspective on Judgment and Choice: Mapping Bounded Rationality,” *American Psychologist* (58:9), pp. 697–720. (https://doi.org/10.1037/0003-066X.58.9.697).

Kahneman, D., and Frederick, S. 2002. “Representativeness Revisited: Attribute Substitution in Intuitive Judgment,” *Heuristics and Biases: The Psychology of Intuitive Judgment* (49), Cambridge University Press, p. 81. (https://doi.org/10.1017/cbo9780511808098.004).

Kasanen, E., and Lukka, K. 1993. “The Constructive Approach in Management Accounting Research,” *Journal of Management Accounting Research* (5:5), Sarasota, pp. 243–264.

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A., and van Bruggen, G. H. 2009. “How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations,” *Information Systems Research*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. 2017. “LightGBM : A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3149–3157.

Keeney, R. L., Raiffa, H., and others. 1993. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, Cambridge university press.

Kleinberg, S., and Hripcsak, G. 2011. “A Review of Causal Inference for Biomedical Informatics,” *Journal of Biomedical Informatics* (44:6), Elsevier, pp. 1102–1112. (https://doi.org/10.1016/j.jbi.2011.07.001).

von Krogh, G. 2018. “Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing,” *Academy of Management Discoveries* (4:4), pp. 404–409.

(https://doi.org/10.5465/amd.2018.0084).

Kruse, C. L., Seidel, S., and Gregor, S. 2015. "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions," in *Proceedings of the Annual Hawaii International Conference on System Sciences* (Vol. 2015-March), pp. 4039–4048. (https://doi.org/10.1109/HICSS.2015.485).

Langseth, H., and Nielsen, T. D. 2003. "Fusion of Domain Knowledge with Data for Structural Learning in Object Oriented Domains," *Journal of Machine Learning Research* (4:Jul), pp. 339–368.

Larkin, J. H., and Simon, H. A. 1987. "Why a Diagram Is (Sometimes) Worth Ten Thousand Words," *Cognitive Science* (11:1), John Wiley & Sons, Ltd (10.1111), pp. 65–100.

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*, Harvard university press.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. 2011. "Big Data, Analytics and the Path from Insights to Value," *MIT Sloan Management Review* (52:2), p. 21.

Lawrence, R., Perlich, C., Rosset, S., Khabibrakhmanov, I., Mahatma, S., Weiss, S., Callahan, M., Collins, M., Ershov, A., and Kumar, S. 2010. "Operations Research Improves Sales Force Productivity at IBM," *Interfaces* (40:1), pp. 33–46.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," *Science* (343:6176), American Association for the Advancement of Science, pp. 1203–1205. (https://doi.org/10.1126/science.1248506).

Lebovitz, S. 2019. "Diagnostic Doubt and Artificial Intelligence: An Inductive Field Study of Radiology Work," in *ICIS 2019 Proceedings*, AIS electronic library.

Leek, J. T., and Peng, R. D. 2015. "What Is the Question? Mistaking the Type of Question Being Considered Is the Most Common Error in Data Analysis," *Science* (347:6228), pp. 1314–1315. (https://doi.org/10.1126/science.aaa6146).

Leonard, J. S., and others. 1954. "The Foundations of Statistics," *NY, John Wiley*, pp. 188–190.

Lerner, J. S., and Tetlock, P. E. 1999. "Accounting for the Effects of Accountability," *Psychological Bulletin* (125:2), American Psychological Association, pp. 255–275. (https://doi.org/10.1037/0033-2909.125.2.255).

Li, T. (Carol), and Chan, Y. E. 2019. "Dynamic Information Technology Capability: Concept Definition and Framework Development," *The Journal of Strategic Information Systems* (28:4), p. 101575. (https://doi.org/10.1016/j.jsis.2019.101575).

Lightfoot, H., Baines, T., and Smart, P. 2013. "The Servitization of Manufacturing: A Systematic Literature Review of Interdependent Trends," *International Journal of Operations & Production Management* (33:11–12), Emerald Group Publishing Limited, pp. 1408–1434.

Lilien, G. L., Rangaswamy, A., Van Bruggen, G. H., and Starke, K. 2004. "DSS Effectiveness in Marketing Resource Allocation Decisions: Reality vs. Perception," *Information Systems Research* (15:3), INFORMS, pp. 216–235. (https://doi.org/10.1287/isre.1040.0026).

Lipton, P. 1990. "Contrastive Explanation," *Royal Institute of Philosophy Supplement* (27), Cambridge University Press, pp. 247–266.

Liu, S., Cui, W., Wu, Y., and Liu, M. 2014. "A Survey on Information Visualization: Recent Advances and Challenges," *The Visual Computer* (30:12), Springer, pp. 1373–1393.

Lundberg, Scott M., Erion, G. G., and Lee, S.-I. 2018. *Consistent Individualized Feature Attribution for Tree Ensembles*.

Lundberg, S. M., and Lee, S.-I. 2017. *Consistent Feature Attribution for Tree Ensembles*.

Lundberg, S. M., and Lee, S. I. 2017. "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, pp. 4766–4775.

Lundberg, Scott M, Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K. W., Newman, S. F., Kim, J., and Lee, S. I. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery," *Nature Biomedical Engineering* (2:10), pp. 749–760. (https://doi.org/10.1038/s41551-018-0304-0).

Luoto, S., Brax, S. A., and Kohtamäki, M. 2017. "Critical Meta-Analysis of Servitization Research: Constructing a Model-Narrative to Reveal Paradigmatic Assumptions," *Industrial Marketing Management* (60), Elsevier, pp. 89–100.

Lycett, M. 2013. "'Datafication': Making Sense of (Big) Data in a Complex World," *European Journal of Information Systems* (22:4), Taylor & Francis, pp. 381–386. (https://doi.org/10.1057/ejis.2013.10).

Lyytinen, K., Nickerson, J. V, and King, J. L. 2020. "Metahuman Systems = Humans + Machines That Learn," *Journal of Information Technology*, p. 0268396220915917. (https://doi.org/10.1177/0268396220915917).

Ma, S.-H., and Liu, J.-L. 2007. "The MCMC Approach for Solving the Pareto/NBD Model and Possible Extensions," in *Third International Conference on Natural Computation (ICNC 2007)* (Vol. 2), pp. 505–512.

Maier, N. R. F. 1963. *Problem-Solving Discussions and Conferences: Leadership Methods and Skills*, McGraw-Hill Series in Management, McGraw-Hill.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. 1982. "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *Journal of Forecasting* (1:2), Wiley Online Library, pp. 111–153.

Mano, H. 1992. "Judgments under Distress: Assessing the Role of Unpleasantness and Arousal in Judgment Formation," *Organizational Behavior and Human Decision Processes* (52:2), Elsevier, pp. 216–245. (https://doi.org/10.1016/0749-5978(92)90036-7).

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., and Sanghvi, S. 2017. "Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation," New York City, NY, USA: McKinsey Global Institute.

Mao, J. 1995. "An Experimental Study of the Use and Effects of Hypertext-Based Explanations in Knowledge-Based Systems," University of British Columbia.

March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), North-Holland, pp. 251–266. (https://doi.org/10.1016/0167-9236(94)00041-2).

Marchand, D. A., and Peppard, J. 2013. "Why IT Fumbles Analytics," *Harvard Business Review* (91:1–2), Harvard Business School Publishing, pp. 104–112.

Markus, M. L. 2017. "Datification, Organizational Strategy, and IS Research: What's the Score?," *The Journal of Strategic Information Systems* (26:3), Elsevier, pp. 233–241.

Markus, M. L., Majchrzak, A., and Gasser, L. 2002. "A Design Theory for Systems That Support Emergent Knowledge Processes," *MIS Quarterly: Management Information Systems* (26:3), JSTOR, pp. 179–212.

Markus, M. L., and Robey, D. 1988. "Information Technology and Organizational Change: Causal Structure in Theory and Research," *Management Science* (34:5), pp. 583–598.

Martens, D., Baesens, B., Van Gestel, T., and Vanthienen, J. 2007. "Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines," *European Journal of Operational Research* (183:3), Elsevier, pp. 1466–1476. (https://doi.org/10.1016/j.ejor.2006.04.051).

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly: Management Information Systems* (38:1), NYU Working Paper, pp. 73–99. (https://doi.org/10.25300/MISQ/2014/38.1.04).

Martínez-López, F. J., and Casillas, J. 2013. "Artificial Intelligence-Based Systems Applied in Industrial Marketing: An Historical Overview, Current and Future Insights," *Industrial Marketing Management* (42:4), Elsevier, pp. 489–495. (https://doi.org/10.1016/j.indmarman.2013.03.001).

Mathiassen, L. 2002. "Collaborative Practice Research," *Information Technology & People* (15:4), MCB UP Ltd, pp. 321–345. (https://doi.org/10.1108/09593840210453115).

McAllister, D. W., Mitchell, T. R., and Beach, L. R. 1979. "The Contingency Model for the Selection of Decision Strategies: An Empirical Test of the Effects of Significance, Accountability, and Reversibility," *Organizational Behavior and Human Performance* (24:2), Elsevier, pp. 228–244. (https://doi.org/10.1016/0030-5073(79)90027-8).

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., and Shetty, S. 2020. "International Evaluation of an AI System for Breast Cancer Screening," *Nature* (577:7788), Nature Publishing Group, pp. 89–94. (https://doi.org/10.1038/s41586-019-1799-6).

Meehl, P. E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.*, University of Minnesota Press.

Merton, R. K., and Merton, R. C. 1968. *Social Theory and Social Structure*, Simon and Schuster.

Miller, G. A. 1956. "The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for

Processing Information.," *Psychological Review* (63:2), American Psychological Association, p. 81.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, Elsevier B.V., pp. 1–38.

Mitchell, T. M., Mabadevan, S., and Steinberg, L. I. 1990. "LEAP: A Learning Apprentice for VLSI Design," in *Machine Learning*, Elsevier, pp. 271–289.

Moffitt, K. E. 1989. *An Empirical Test of Expert System Explanation Facility Effects on Incidental Learning and Decision-Making*, Arizona State University.

Monat, J. P. 2011. "Industrial Sales Lead Conversion Modeling," *Marketing Intelligence & Planning* (29:2), Emerald Group Publishing Limited, pp. 178–194.

Mora Cortez, R., and Johnston, W. J. 2017. "The Future of B2B Marketing Theory: A Historical and Prospective Analysis," *Industrial Marketing Management* (66), Elsevier Inc., pp. 90–102. (https://doi.org/10.1016/j.indmarman.2017.07.017).

Morgan, S. L., and Winship, C. 2015. *Counterfactuals and Causal Inference*, Cambridge University Press.

Müller, O., Fay, M., and vom Brocke, J. 2018. "The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics," *Journal of Management Information Systems* (35:2), Routledge, pp. 488–509. (https://doi.org/10.1080/07421222.2018.1451955).

Müller, O., Junglas, I., Brocke, J. vom, and Debortoli, S. 2016. "Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines," *European Journal of Information Systems* (25:4), Taylor & Francis, pp. 289–302. (https://doi.org/10.1057/ejis.2016.2).

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*.

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification,'" *The Journal of Strategic Information Systems* (24:1), pp. 3–14. (https://doi.org/10.1016/j.jsis.2015.02.001).

Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), SAGE PublicationsSage CA: Los Angeles, CA, pp. 175–220.

Nilsson, N. J. 1998. *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann.

Orlikowski, W. J., and Iacono, C. S. 2001. "Research Commentary: Desperately Seeking the 'IT' in IT Research - A Call to Theorizing the IT Artifact," *Information Systems Research*. (https://doi.org/10.1287/isre.12.2.121.9700).

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. 2000. "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* (30:3), pp. 286–297. (https://doi.org/10.1109/3468.844354).

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. 2015. "Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques," *Expert Systems with Applications* (42:1), Elsevier, pp. 259–268. (https://doi.org/10.1016/j.eswa.2014.07.040).

Pearl, J. 1995. “Causal Diagrams for Empirical Research,” *Biometrika* (82:4), Oxford University Press, p. 669. (https://doi.org/10.2307/2337329).

Pearl, J. 1999. “Probabilities of Causation: Three Counterfactual Interpretations and Their Identification,” *Synthese* (121:1–2), Springer, pp. 93–149.

Pearl, J. 2009. *Causality*, Cambridge university press.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why : The New Science of Cause and Effect*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems* (24:3), Taylor & Francis, pp. 45–77.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz Harvard, B., Lyons, T., Manyika, J., Carlos Niebles, J., and Mishra, S. 2019. *“The AI Index 2019 Annual Report”, AI Index Steering Committee, Human-Centered AI Institute*, Stanford University, Stanford, CA.

Peter E. Rossi, author, and Greg M. Allenby, author. 2003. “Bayesian Statistics and Marketing,” *Marketing Science* (3), p. 304.

Plastino, E., and Purdy, M. 2018. “Game Changing Value from Artificial Intelligence: Eight Strategies,” *Strategy & Leadership* (46:1), pp. 16–22. (https://doi.org/10.1108/SL-11-2017-0106).

Platzer, M., and Reutterer, T. 2016. “Ticking Away the Moments: Timing Regularity Helps to Better Predict Customer Activity,” *Marketing Science* (35:5), pp. 779–799. (https://doi.org/10.1287/mksc.2015.0963).

Popper, K. R. 2002. *An Unended Quest*, Psychology Press.

Price, P. C., Jhangiani, R. S., and Chiang, I.-C. A. 2015. “Quasi-Experimental Research,” *Research Methods in Psychology*.

Purao, S., Rossi, M., and Sein, M. K. 2010. *On Integrating Action Research and Design Research*, pp. 179–194. (https://doi.org/10.1007/978-1-4419-5653-8_13).

Rai, A., Constantinides, P., and Sarker, S. 2019. “Editor’s Comments: Next-Generation Digital Platforms: Toward Human–AI Hybrids,” *Management Information Systems Quarterly* (43:1), iii–ix.

Raisch, S., and Krakowski, S. 2020. “Artificial Intelligence and Management: The Automation-Augmentation Paradox,” *Academy of Management Review*. (https://doi.org/10.5465/2018.0072).

Ransbotham, S., Kiron, D., and Prentice, P. K. 2015. “The Talent Dividend,” *MIT Sloan Management Review* (56:4), p. 1.

Remenyi, D., and Sherwood-Smith, M. 1999. “Maximise Information Systems Value by Continuous Participative Evaluation,” *Logistics Information Management* (12:1/2), MCB UP Ltd, pp. 14–31. (https://doi.org/10.1108/09576059910256222).

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016a. “‘Why Should i Trust You?’ Explaining the Predictions of Any Classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 13-17-Augu), pp. 1135–1144.

(https://doi.org/10.1145/2939672.2939778).

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016b. *Model-Agnostic Interpretability of Machine Learning*.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016c. "'Why Should I Trust You?,'" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 1135–1144.

Rich, E., and Knight, K. 1990. *Artificial Intelligence*, (2nd ed.), McGraw-Hill Higher Education.

Robnik-Šikonja, M., and Kononenko, I. 2008. "Explaining Classifications for Individual Instances," *IEEE Transactions on Knowledge and Data Engineering* (20:5), IEEE, pp. 589–600. (https://doi.org/10.1109/TKDE.2007.190734).

Rosemann, M., and Vessey, I. 2008. "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *Mis Quarterly*, JSTOR, pp. 1–22.

Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* (70:1), Oxford University Press, pp. 41–55.

Rothwell, P. M. 2006. "Factors That Can Affect the External Validity of Randomised Controlled Trials," *PLoS Clinical Trials* (1:1), Public Library of Science, p. e9. (https://doi.org/10.1371/journal.pctr.0010009).

Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* (66:5), American Psychological Association, pp. 688–701. (https://doi.org/10.1037/h0037350).

Rubin, D. B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association* (100:469), Taylor & Francis, pp. 322–331.

Russell, S. J., and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*, Malaysia: Pearson education.

Rust, R. T., and Huang, M. H. 2014. "The Service Revolution and the Transformation of Marketing Science," *Marketing Science* (33:2), INFORMS, pp. 206–221. (https://doi.org/10.1287/mksc.2013.0836).

SAS. 2012. "The Evolution of Decision Making: How Leading Organizations Are Developing a Data-Driven Culture."

Schisterman, E. F., Perkins, N. J., Mumford, S. L., Ahrens, K. A., and Mitchell, E. M. 2017. "Collinearity and Causal Diagrams: A Lesson on the Importance of Model Specification Enrique," *Epidemiology* (28:1), NIH Public Access, pp. 47–53. (https://doi.org/10.1097/EDE.0000000000000554).

Schmittlein, D. C., Morrison, D. G., and Colombo, R. 1987. "Counting Your Customers: Who-Are They and What Will They Do Next?," *Management Science* (33:1), INFORMS, pp. 1–24.

Van De Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., and Van Loey, N. E. 2015. "Analyzing Small Data Sets Using Bayesian Estimation: The Case of Posttraumatic Stress Symptoms Following Mechanical Ventilation in Burn Survivors," *European Journal of Psychotraumatology*.

(https://doi.org/10.3402/ejpt.v6.25216).

Schultze, U., and Avital, M. 2011. "Designing Interviews to Generate Rich Data for Information Systems Research," *Information and Organization* (21:1), Elsevier, pp. 1–16.

Schulz, K. F., Chalmers, I., Hayes, R. J., and Altman, D. G. 1995. "Empirical Evidence of Bias: Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials," *JAMA: The Journal of the American Medical Association* (273:5), American Medical Association, pp. 408–412. (https://doi.org/10.1001/jama.273.5.408).

Scott, S. L., and Varian, H. 2013. "Bayesian Variable Selection for Nowcasting Economic Time Series," *Economics of Digitization.* (July 2012), pp. 1–22.

Scott, S. L., and Varian, H. R. 2014. "Predicting the Present with Bayesian Structural Time Series," *International Journal of Mathematical Modelling and Numerical Optimisation* (5:1–2), pp. 4–23. (https://doi.org/10.1504/IJMMNO.2014.059942).

Seidel, S., Chandra Kruse, L., Székely, N., Gau, M., and Stieger, D. 2018. "Design Principles for Sensemaking Support Systems in Environmental Sustainability Transformations," *European Journal of Information Systems*. (https://doi.org/10.1057/s41303-017-0039-0).

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R. 2011. "Action Design Research," *MIS Quarterly: Management Information Systems* (35:1), pp. 37–56. (https://doi.org/10.2307/23043488).

Shadish, W. R., Cook, T. D., Campbell, D. T., and others. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell.*, Boston: Houghton Mifflin,.

Shaikhina, T., and Khovanova, N. A. 2017. "Handling Limited Datasets with Neural Networks in Medical Applications: A Small-Data Approach," *Artificial Intelligence in Medicine* (75), pp. 51–63. (https://doi.org/10.1016/j.artmed.2016.12.003).

Sharma, A. 1997. "Professional as Agent: Knowledge Asymmetry in Agency Exchange," *Academy of Management Review* (22:3), Academy of Management Briarcliff Manor, NY 10510, pp. 758–798. (https://doi.org/10.5465/AMR.1997.9708210725).

Sharma, R., Mithas, S., and Kankanhalli, A. 2014. "Transforming Decision-Making Processes: A Research Agenda for Understanding the Impact of Business Analytics on Organisations," *European Journal of Information Systems* (Vol. 23), Taylor & Francis, pp. 433–441.

Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., and Herdlein, S. a. 2000. "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing*.

Sheridan, T. B., and Verplank, W. L. 1978. "Human and Computer Control of Undersea Teleoperators," Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. 2002. "Past, Present, and Future of Decision Support Technology," *Decision Support Systems* (33:2), Elsevier, pp. 111–126. (https://doi.org/10.1016/S0167-9236(01)00139-7).

Shirley, G., Oliver, M., and Stefan, S. 2013. “Reflection, Abstraction, and Theorizing in Design and Development Research,” *ECIS 2013 - Proceedings of the 21st European Conference on Information Systems*.

Shmueli, G. 2010. “To Explain or to Predict?,” *Statistical Science* (25:3), Institute of Mathematical Statistics, pp. 289–310.

Shmueli, G., and Koppius, O. R. 2011. “Predictive Analytics in Information Systems Research,” *MIS Quarterly*, JSTOR, pp. 553–572.

Shollo, A., and Galliers, R. D. 2016. “Towards an Understanding of the Role of Business Intelligence Systems in Organisational Knowing,” *Information Systems Journal* (26:4), Wiley Online Library, pp. 339–367.

Shrestha, Y. R., Ben-Menahem, S. M., and von Krogh, G. 2019. “Organizational Decision-Making Structures in the Age of Artificial Intelligence,” *California Management Review* (61:4), SAGE Publications Ltd, pp. 66–83. (https://doi.org/10.1177/0008125619862257).

Siegel-Jacobs, K., and Yates, J. F. 1996. “Effects of Procedural and Outcome Accountability on Judgment Quality,” *Organizational Behavior and Human Decision Processes* (65:1), Elsevier, pp. 1–17. (https://doi.org/10.1006/obhd.1996.0001).

Silver, N. 2012. *The Signal and the Noise: Why so Many Predictions Fail--but Some Don’t*, Penguin.

Simon, H. A. 1955. “A Behavioral Model of Rational Choice,” *The Quarterly Journal of Economics* (69:1), pp. 99–118. (https://doi.org/10.2307/1884852).

Simon, H. A. 1956. “Rational Choice and the Structure of the Environment.,” *Psychological Review* (63:2), American Psychological Association, p. 129.

Simon, H. A. 1979. “Rational Decision Making in Business Organizations,” *American Economic Review* (69:4), American Economic Association, pp. 493–513. (https://doi.org/10.2307/1808698).

Simon, H. A. 1997. *Administrative Behavior*, Simon and Schuster.

Simon, H. A. (Herbert A. 2019. *The Sciences of the Artificial*, MIT Press.

Simonson, I., Staw, B. M., and Haas, W. A. 1992. “Deescalation Strategies: A Comparison of Techniques for Reducing Commitment to Losing Courses of Action,” *Journal of Applied Psychology* (Vol. 77).

Sinha, A. P., and Zhao, H. 2008. “Incorporating Domain Knowledge into Data Mining Classifiers: An Application in Indirect Lending,” *Decision Support Systems* (46:1), Elsevier, pp. 287–299.

Sodenkamp, M., Kozlovskiy, I., and Staake, T. 2015. “Gaining IS Business Value through Big Data Analytics: A Case Study of the Energy Sector,” *ICIS 2015 Proceedings*.

Stanovich, K. E., and West, R. F. 2003. “"Individual Differences in Reasoning: Implications for the Rationality Debate?,” *Behavioral and Brain Sciences* (26:4), Cambridge University Press, p. 527. (https://doi.org/10.1017/S0140525X03210116).

Stormi, K., Laine, T., and Elomaa, T. 2018. “Feasibility of B2C Customer Relationship Analytics in the B2B Industrial Context,” *26th European Conference on Information Systems: Beyond Digitization -*

Facets of Socio-Technical Change, ECIS 2018.

Štrumbelj, E., and Kononenko, I. 2010. "An Efficient Explanation of Individual Classifications Using Game Theory," *Journal of Machine Learning Research* (11), JMLR. org, pp. 1–18.

Štrumbelj, E., and Kononenko, I. 2011. *A General Method for Visualizing and Explaining Black-Box Regression Models*, Springer, Berlin, Heidelberg, pp. 21–30.

Štrumbelj, E., and Kononenko, I. 2014. "Explaining Prediction Models and Individual Predictions with Feature Contributions," *Knowledge and Information Systems* (41:3), Springer London, pp. 647–665.

Štrumbelj, E., Kononenko, I., and Robnik Šikonja, M. 2009. "Explaining Instance Classifications with Interactions of Subsets of Feature Values," *Data and Knowledge Engineering* (68:10), Elsevier, pp. 886–904. (https://doi.org/10.1016/j.datak.2009.01.004).

Suchman, L. 2002. "Located Accountabilities in Technology Production," *Scandinavian Journal of Information Systems* (14:2), p. 7.

Sundin, E. 2009. "Life-Cycle Perspectives of Product/Service-Systems: In Design Theory," *Introduction to Product/Service-System Design*, Springer, pp. 31–49. (https://doi.org/10.1007/978-1-84882-909-1_2).

Susman, G. I., and Evered, R. D. 1978. "An Assessment of the Scientific Merits of Action Research," *Administrative Science Quarterly* (23:4), p. 582. (https://doi.org/10.2307/2392581).

Sutanto, J., Kankanhalli, A., Tay, J., Raman, K. S., and Tan, B. C. Y. 2009. "Change Management in Interorganizational Systems for the Public," *Journal of Management Information Systems* (25:3), pp. 133–176.

Talagala, N. 2018. "Operational Machine Learning: Seven Considerations for Successful MLOps." (, accessed August 12, 2020).

Tambe, P. 2014. "Big Data Investment, Skills, and Firm Value," *Management Science* (60:6), INFORMS, pp. 1452–1469. (https://doi.org/10.1287/mnsc.2014.1899).

Tarafdar, M., Beath, C. M., and Ross, J. W. 2019. "Using AI to Enhance Business Operations," *MIT Sloan Management Review* (11), Massachusetts Institute of Technology, Cambridge, MA.

Tashman, L. J. 2000. "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review," *International Journal of Forecasting* (16:4), Elsevier, pp. 437–450.

Tetlock, P. E. 1983. "Accountability and the Perseverance of First Impressions," *Social Psychology Quarterly* (46:4), JSTOR, p. 285. (https://doi.org/10.2307/3033716).

Tetlock, Philip E. 1985. "Accountability: The Neglected Social Context of Judgment and Choice," *Research in Organizational Behavior* (7:1), Greenwich, CT, pp. 297–332.

Tetlock, Philip E. 1985. "Accountability: A Social Check on the Fundamental Attribution Error," *Social Psychology Quarterly* (48:3), JSTOR, p. 227. (https://doi.org/10.2307/3033683).

Tetlock, P. E., Skitka, L., and Boettger, R. 1989. "Social and Cognitive Strategies for Coping with Accountability: Conformity, Complexity, and Bolstering.," *Journal of Personality and Social*

Psychology (57:4), American Psychological Association, p. 632.

Tetlock, P. E., Vieider, F. M., Patil, S. V., and Grant, A. M. 2013. "Accountability and Ideology: When Left Looks Right and Right Looks Left," *Organizational Behavior and Human Decision Processes* (122:1), Academic Press, pp. 22–35. (https://doi.org/10.1016/J.OBHDP.2013.03.007).

Thaler, R. 1980. "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization* (1:1), Citeseer, pp. 39–60. (https://doi.org/10.1016/0167-2681(80)90051-7).

Thaler, R. H., and Sunstein, C. R. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin.

Thiess, T., and Müller, O. 2018. "Towards Design Principles for Data-Driven Decision Making – An Action Design Research Project in the Maritime Industry," *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018*, Portsmouth, UK: European Conference on Information Systems (ECIS).

Thiess, T., and Müller, O. 2020a. "Setting Sail for Data-Driven Decision-Making an Action Design Research Case from the Maritime Industry," in *Design Science Research. Cases*, pp. 291–317. (https://doi.org/10.1007/978-3-030-46781-4_12).

Thiess, T., and Müller, O. 2020b. "Designing Causal Inference Systems for Value-Based Spare Parts Pricing: An ADR Study at MAN Energy Solutions," *Lecture Notes in Business Information Processing* (Vol. 398 LNBIP). (https://doi.org/10.1007/978-3-030-61140-8_13).

Thiess, T., Müller, O., and Tonelli, L. 2020. "Design Principles for Explainable Sales Win-Propensity Prediction Systems," *WI2020 Zentrale Tracks*, GITO Verlag, pp. 326–340. (https://doi.org/10.30844/wi_2020_c8-thiess).

Totte Harinen, and Bonnie Li. 2019. "Using Causal Inference to Improve the Uber User Experience," *Uber Engineering*, Jun, pp. 1–11.

Toulmin, S. 1958. "The Uses of Argument Cambridge University Press," *Cambridge, UK*.

Tversky, A., and Kahneman, D. 1974. "Judgment under Uncertainty: Heuristics and Biases," *Science* (185:4157), pp. 1124–1131. (https://doi.org/10.1126/science.185.4157.1124).

Tversky, A., and Kahneman, D. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty* (5:4), Springer, pp. 297–323. (https://doi.org/10.1007/BF00122574).

Umanath, N. S., and Vessey, I. 1994. "Multiattribute Data Presentation and Human Judgment: A Cognitive Fit Perspective," *Decision Sciences* (25:5–6), Wiley Online Library, pp. 795–824. (https://doi.org/10.1111/j.1540-5915.1994.tb01870.x).

Varian, H. R. 2014. "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives* (28:2), American Economic Association, pp. 3–28. (https://doi.org/10.1257/jep.28.2.3).

Varian, H. R. 2016. "Causal Inference in Economics and Marketing," *Proceedings of the National Academy of Sciences of the United States of America* (113:27), National Acad Sciences, pp. 7310–7315. (https://doi.org/10.1073/pnas.1510479113).

Wagner, M. M., and Hogan, W. R. 1996. “The Accuracy of Medication Data in an Outpatient Electronic Medical Record,” *Emerging Infectious Diseases* (3:3), pp. 234–244. (https://doi.org/10.1136/jamia.1996.96310637).

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. “Building an Information System Design Theory for Vigilant EIS,” *Information Systems Research* (3:1), INFORMS , pp. 36–59. (https://doi.org/10.1287/isre.3.1.36).

Watson, H. J. 2014. “Tutorial: Big Data Analytics: Concepts, Technologies, and Applications,” *Communications of the Association for Information Systems* (34), pp. 1247–1268. (https://doi.org/10.17705/1CAIS.03462).

Weizenbaum, J. 1983. “ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine,” *Communications of the ACM* (26:1), ACM New York, NY, USA, pp. 23–28. (https://doi.org/10.1145/357980.357991).

Wheelwright, S., Makridakis, S., and Hyndman, R. J. 1998. *Forecasting: Methods and Applications*, John Wiley & Sons.

Wickboldt, C., and Kliewer, N. 2018. “Value Based Pricing Meets Data Science: A Concept for Automated Spare Part Valuation.”

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., and Sarter, N. B. 2010. “Stages and Levels of Automation: An Integrated Meta-Analysis,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54), Sage Publications Sage CA: Los Angeles, CA, pp. 389–393.

Winner, L. 1980. “Do Artifacts Have Politics?,” *Daedalus*, JSTOR, pp. 121–136.

Winston, P. H. 1992. “Artificial Intelligence 3rd Edition,” *Addison-Wesley, Reading, MA* (34), pp. 167–339.

Wiseman, R. M., and Gomez-Mejia, L. R. 1998. “A Behavioral Agency Model of Managerial Risk Taking,” *Academy of Management Review* (23:1), Academy of Management Briarcliff Manor, NY 10510, pp. 133–153. (https://doi.org/10.5465/AMR.1998.192967).

Wolpert, D. H., and Macready, W. G. 1996. “No Free Lunch Theorems for Optimization.”

Wu, L., Hitt, L., and Lou, B. 2019. “Data Analytics, Innovation, and Firm Productivity,” *Management Science* (66:5), pp. 2017–2039. (https://doi.org/10.1287/mnsc.2018.3281).

Wu, L., and Hitt, L. M. 2016. *How Do Data Skills Affect Firm Productivity: Evidence from Process-Driven vs. Innovation-Driven Practices*.

Wu, L., Hitt, L. M., and Lou, B. 2017. *Data Analytics Skills, Innovation and Firm Productivity*.

Xu, X., Tang, L., and Rangan, V. 2017. “Hitting Your Number or Not? A Robust & Intelligent Sales Forecast System,” in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, December, pp. 3613–3622.

Yan, J., Gong, M., Sun, C., Huang, J., and Chu, S. M. 2015. “Sales Pipeline Win Propensity Prediction: A Regression Approach,” in *2015 IFIP/IEEE International Symposium on Integrated Network*

Management (IM), IEEE, May, pp. 854–857. (https://doi.org/10.1109/INM.2015.7140393).

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S., and Yang, X. 2015. "On Machine Learning towards Predictive Sales Pipeline Analytics," *Proceedings of the National Conference on Artificial Intelligence* (Vol. 3), , February 18.

Ye, L. R. 1991. *User Requirements for Explanation in Expert Systems.*

Zhang, C., Li, X., Yan, J., Qui, S., Wang, Y., Tian, C., and Zhao, Y. 2014. "Sufficient Statistics Feature Mapping over Deep Boltzmann Machine for Detection," in *2014 22nd International Conference on Pattern Recognition*, IEEE, August, pp. 827–832.

Zotteri, G., and Kalchschmidt, M. 2007. "A Model for Selecting the Appropriate Level of Aggregation in Forecasting Processes," *International Journal of Production Economics* (108:1–2), Elsevier, pp. 74–83. (https://doi.org/10.1016/j.ijpe.2006.12.030).

Zuboff, S. 1985. "Automate/Informate: The Two Faces of Intelligent Technology," *Organizational Dynamics* (14:2), pp. 5–18.