Streamlining the Operation of AI Systems: Examining MLOps Maturity at an Automotive Firm

Michael Weber
Technical University of Munich
mic.weber@tum.de

Johannes Schniertshauer
AUDI AG
johannes.schniertshauer@audi.de

Leonard Przybilla
Technical University of Munich
leonard.przybilla@tum.de

Andreas Hein
Technical University of Munich
andreas.hein@tum.de

Jörg Weking
Technical University of Munich
joerg.weking@tum.de

Helmut Krcmar
Technical University of Munich
helmut.krcmar@tum.de

Abstract

Developing and operating AI systems based on machine learning (ML) has unique challenges that render traditional practices inappropriate (e.g., managing data drift). To that end, MLOps emerged as a novel paradigm for managers and teams to develop and operate such ML systems successfully. Organizations currently employ different maturity levels for MLOps, whereas higher maturity typically corresponds to more automated, streamlined, and reliable workflows. However, we have limited insight into factors influencing MLOps maturity in ML projects. Therefore, we conducted a case study on MLOps maturity in three ML projects at an automotive firm. We identified several contextual factors that facilitate or inhibit MLOps maturity, such as the ML model's complexity, the quality of new data, and the appropriateness of available MLOps tools. Our study contributes to research on managing and organizing AI by providing factors that explain the different adoption of MLOps in practice.

Keywords: Machine Learning, MLOps, Artificial Intelligence, Operation, Deployment.

1. Introduction

Artificial intelligence (AI) presents organizations with unprecedented capabilities to automate tasks, augment human capabilities, and develop new products or services (Böttcher et al., 2022; Sjödin et al., 2021). Much of the recent developments in AI can be attributed to advances in machine learning (ML), the vast availability of data, and increasing computing power (Russell & Norvig, 2021). ML enables organizations to build systems that can learn from experience and act increasingly autonomously while becoming more inscrutable to human developers and users (Berente et al., 2021). However, due to those novel characteristics, ML systems pose several challenges to management and organizing (Berente et al., 2021; Lyytinen et al., 2021), some of which relate to the development and operations of ML systems.

Indeed, while ML development and operations are essential to creating value from ML systems (Shollo et al., 2022; Sjödin et al., 2021), several surveys have indicated a noticeable gap between ML development and operations (Benbya et al., 2020). As a possible explanation, recent findings suggest that conventional methods for software development do not seamlessly align with the specific requirements for developing and operating ML systems (Dolata et al., 2022; Laato et al., 2022). Challenges include handling data drift, employing automated feedback loops to retrain ML systems and ensuring a continuous account of legal and ethical issues (Lwakatare et al., 2019; Paleyes et al., 2022).

Against this backdrop, MLOps has emerged as a promising paradigm that attempts to address the gap between ML development and operations (Kreuzberger et al., 2023; Ruf et al., 2021; Testi et al., 2022). In short, MLOps extends the DevOps method with workflows and principles tailored to the specific characteristics of ML systems (CD Foundation, 2022). Depending on how organizations employ these workflows and principles, they can achieve different maturity levels of MLOps (John et al., 2021; Ruf et al., 2021). Higher MLOps maturity is typically associated with more automated, streamlined, and reliable ML workflows (cf. Google Cloud, 2020; John et al., 2021; Microsoft, 2023). Therefore, a higher MLOps maturity is generally desirable for organizations.

However, we lack detailed insight into which factors influence the different levels of MLOps maturity observable in practice. To the best of our

URI: https://hdl.handle.net/10125/107091
978-0-9981331-7-1
(CC BY-NC-ND 4.0)

knowledge, only a few studies in Information Systems (IS) have considered the operational phase of ML systems in detail (e.g., Grønsund & Aanestad, 2020; Waardenburg & Huysman, 2022). Moreover, there is thus far only anecdotal evidence of how organizations employed MLOps workflows and principles (e.g., Granlund, Stirbu, et al., 2021; Laato et al., 2022). Thus, we ask the research question: *Which factors influence MLOps maturity in ML projects?*

To address this question, we conduct a case study (Yin, 2018) on MLOps maturity with three ML projects at AutoCorp, an automotive firm. This case study aims to identify relevant factors influencing MLOps maturity in organizational practice. We conducted 17 semi-structured interviews and relied on established methods for qualitative data analysis to derive our findings (Miles & Huberman, 1994). Drawing on our analysis, we derive and discuss factors influencing MLOps maturity in ML projects.

2. Background

2.1 DevOps

DevOps (Development and Operations) builds on agile software development and essentially refers to a **method** that aims to shorten the release times for software updates (Fitzgerald & Stol, 2017). Thereby, DevOps attempts to address the gap and misalignment between software development and operations (Fitzgerald & Stol, 2017). Essential principles of this method are **continuous integration and continuous deployment (CI/CD)**, i.e., the continuous integration and delivery of a software application by a team of developers. In addition, DevOps emphasizes interdisciplinary collaboration, the autonomy of teams, knowledge sharing, communication, tool-supported automation, and efficiency in workflows (Gall & Pigni, 2022; Wiedemann et al., 2019).

IS research has shown how DevOps practices shape intra-organizational IT alignment (Wiedemann et al., 2020) and influence job satisfaction (Hemon-Hildgen et al., 2020). However, adopting DevOps is also bound to several challenges, including upskilling requirements, lack of communication, cultural barriers, feasibility, and environmental factors (Leite et al., 2019; Riungu-Kalliosaari et al., 2016).

2.2 MLOps: Workflow and Maturity Levels

The Continuous Delivery Foundation's Special Interest Group MLOps defines MLOps as "**the extension of the DevOps methodology** to **include Machine Learning and Data Science assets** as first-class citizens within the DevOps ecology" (CD Foundation, 2022). Extending traditional DevOps to include ML specifics seems necessary due to the unique requirements of developing and operating ML systems (CD Foundation, 2022; Google Cloud, 2020; Symeonidis et al., 2022). These ML-specific requirements include, among many others, the need for continuous monitoring of ML systems and much experimentation involved in their development. Figure 1 illustrates a typical MLOps workflow.

As with DevOps, MLOps is not about merely developing and operating a software artifact; instead, MLOps focuses on *how* teams can achieve this more efficiently and reliably (CD Foundation, 2022). Literature highlights several essential principles for MLOps (Kreuzberger et al., 2023; Ruf et al., 2021; Testi et al., 2022). We synthesize these principles in three areas. First, the **continuous training, integration, and deployment** of ML systems are essential principles that enhance the scope of traditional CI/CD (e.g., with the continuous (re-)training of ML systems). Teams use various MLOps tools to automate the ML pipeline from data extraction to deployment. Second, **continuous monitoring and feedback loops** are novel principles that help prevent ML systems' degrading performance. In practice, teams employ tool- or human-supported monitoring of input data and the ML model's performance to (semi-

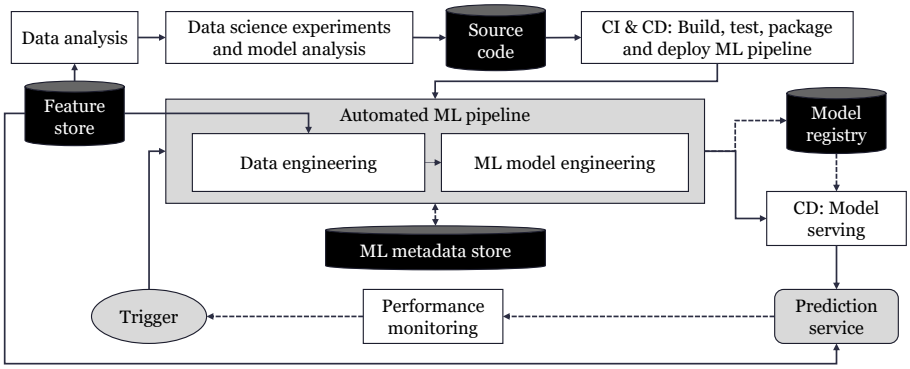

Figure 1. MLOps workflow (adapted from Google Cloud (2020)))

) automatically trigger an update of the ML system. The updating typically involves a retraining of the underlying ML model. Third, **traceability and reproducibility** are other principles that help make ML development and operations more transparent and allow teams to reproduce specific workflow steps (e.g., data science experiments). In practice, teams employ tools to version data and ML models and to track relevant meta-data (e.g., hyperparameters).

The workflow and principles of MLOps can be employed differently across ML projects (Laato et al., 2022; Ruf et al., 2021). Depending on the **level of workflow automation** and **adherence to MLOps principles**, research and practice have proposed different levels of **MLOps maturity** (Google Cloud, 2020; John et al., 2021; Microsoft, 2023). We synthesize these ideas in Table 1. We view MLOps maturity as a continuum from manual to fully automated and transparent ML workflows. Higher MLOps maturity helps to reduce manual efforts and deployment times and can contribute to more reliable ML systems. Therefore, a higher MLOps maturity is generally desirable for organizations.

Table 1. MLOps maturity (synthesis from Google Cloud (2020); John et al. (2021); Microsoft (2023))

Maturity	Characteristics
Low	- Manual training, testing, and deployment of the ML system - No monitoring of the ML system - Lack of transparency and reproducibility of ML workflow
Moderate	- Automated (re-)training (with potentially manual deployment steps) - Manual monitoring of the ML system - Transparent and reproducible ML workflow
High	- Fully automated and flexible ML pipeline (including deployment) - Automated and advanced monitoring (e.g., drift detection) - Transparent and reproducible ML workflow

2.3. Related Work

An emergent stream of literature has started looking into the **MLOps** paradigm to address ML-specific challenges and to streamline ML development and operation. John et al. (2021) report on three cases of MLOps with mixed MLOps maturity that, however, all strive for higher MLOps maturity. This is similar to Lwakatare et al. (2019) and Mucha et al. (2022) who argue that organizations tend to gradually evolve toward more mature ML practices. More efficient and reliable practices are especially necessary when operating critical ML systems, such as those acting without human intervention (Lwakatare et al., 2019).

There are several challenges to applying MLOps, which could affect MLOps maturity in practice. One challenge of MLOps is building a stable and repeatable **ML pipeline** (Hill et al., 2016; Martínez-Fernández et al., 2022). The market for MLOps tools is still evolving and heterogeneous (AI Infrastructure Alliance, 2022), which can complicate stable integration (Ruf et al., 2021). Furthermore, a lack of specific tools can complicate MLOps and result in the development of custom tooling (Ackermann et al., 2018; Hill et al., 2016).

Regarding data, the manual effort for maintaining a sufficient data quality and lack of access to experts for data labeling tasks can create a bottleneck for retraining ML models (Paleyes et al., 2022; Ruf et al., 2021; Symeonidis et al., 2022). This becomes especially critical when new training data is frequently needed and the ML team has limited control over the data collection process (Nahar et al., 2022). Moreover, multi-organizational settings can further complicate data sharing and integration (Granlund, Kopponen, et al., 2021; Lwakatare et al., 2019; Nahar et al., 2022).

Another challenge relates to the appropriate verification of ML systems on a continuous basis, which is oftentimes not formalized, hard to reliably conduct, and often involves much manual work (Martínez-Fernández et al., 2022; Nahar et al., 2022; Paleyes et al., 2022). In addition, debugging and error tracing of ML systems can become complicated with opaque deep learning systems and hidden feedback loops (Martínez-Fernández et al., 2022).

Last, organizational aspects such as team structure and lack of relevant expertise can complicate MLOps (Kreuzberger et al., 2023; Martínez-Fernández et al., 2022; Nahar et al., 2022). For example, teams without dedicated DevOps or MLOps specialists might require pure data scientists to setup ML pipelines and related infrastructure, which they might lack experience with.

To the best of our knowledge, while research has provided important early empirical insight on MLOps workflows and principles, we thus far lack a detailed account of the factors influencing MLOps maturity at different organizations. However, such an investigation would advance our conceptual understanding of MLOps as a novel paradigm and provide valuable guidance to practitioners.

4. Research Method

We conducted a case study (Yin, 2018) in an automotive firm (AutoCorp), where we examined

MLOps maturity in three ML projects. A case study approach is particularly suited as MLOps presents a novel phenomenon. AutoCorp was selected as a revelatory case site, as AutoCorp allowed us to closely examine the employment of MLOps in three ML projects to distill potentially influencing factors for MLOps maturity. Next, we introduce the case study's setting and depict our data collection and analyses.

4.1 Case Study Setting

AutoCorp is an established original equipment manufacturer in the automotive industry. AutoCorp is strongly pushing the digital transformation of production and logistics to enable more efficient and resilient operations. A key pillar of AutoCorp's digital transformation strategy is the use of data for data analytics and ML systems. Like many other large and established firms, AutoCorp is currently in a phase where the first successful ML systems are introduced to production to create initial value. Nevertheless, AutoCorp still faces various challenges for ML operations that adopting MLOps workflows and principles could address. Therefore, AutoCorp presents a suitable site to study the factors influencing MLOps maturity and was selected as a revelatory case.

4.2 Data Collection

We started our data inquiry with a first round of data collection in September 2021. The first round aimed to understand the general state of ML development and operations at AutoCorp and the context of the firm. We conducted ten semi-structured interviews (Myers & Newman, 2007) with various stakeholders from business, IT, ML projects, and innovation labs. Following the initial analysis, various challenges surrounding ML operations, such as the continued monitoring of ML systems, emerged as an essential barrier to ML adoption. In addition, our data indicated that ML projects at AutoCorp were at different stages of maturity regarding MLOps.

Based on those observations, we engaged in a second round of data collection from October 2022. This round aimed to examine MLOps maturity and potential influencing factors in three specific ML projects at AutoCorp. We sampled these ML projects because of their advanced state and strong interest in MLOps workflows and principles. Furthermore, our sampling aimed at covering different ML types, data sources, and deployment settings (cf. Table 2). We proceeded with data collection with seven interviews with stakeholders from IT and the selected ML projects to examine their MLOps maturity.

We used our conceptual preunderstanding of MLOps workflows, principles, and maturity levels to guide our second round of interviews. For example, we asked about the state of the ML pipeline, the current use of MLOps tools, feedback loops, and open challenges. In addition, we received regular updates on the ML projects during biweekly meetings as part of the joint research project. The guidelines of AutoCorp did not allow us to record the interviews. Therefore, we took detailed notes during the interviews. We made sure that at least two researchers regularly took part in the interviews so that one could solely observe and take detailed notes. If we were uncertain about specific statements, we used our access to AutoCorp to ask follow-up questions for clarification informally. Table 3 summarizes our two rounds of data collection.

Table 2. ML projects examined

ID	Description
ML-P1	Supervised deep learning on image data for error detection (e.g., cracks)
ML-P2	Supervised deep learning on image data for error detection (e.g., scratches)
ML-P3	Supervised shallow machine learning on structured sensor data for error classification (e.g., weld spatter)

Table 3. Data collection

Round 1: Understand ML development and operations at AutoCorp	
Timeline	Sep 21 – Nov 21
Primary sources	10 Interviews* (30-45 mins): - 2x ML engineers - ML platform manager - 2x Business stakeholders - IT manager - IT project manager - IT architect - 2x Innovation engineers
Informal sources	Biweekly project meetings (follow-up questions and clarifications)
Round 2: Examine MLOps maturity in three ML projects at AutoCorp	
Timeline	Oct 22 – Nov 22
Primary sources	7 Interviews* (45 mins): - 4x ML engineers - Head of data science - ML platform manager - IT project manager
Informal sources	Biweekly project meetings (follow-up questions and clarifications)
* The research team did not record interviews due to confidentiality, but we relied on extensive notes	

4.3 Data Analysis

We analyzed our interview notes using qualitative data analysis (Miles & Huberman, 1994). We initially coded words and passages related to MLOps to reduce data, including the applied procedures, tools, and challenges and planned the next steps. In addition, we coded contextual information, such as the current deployment status and organizational structure of an ML project. We further marked whether the codes related to specific ML projects or AutoCorp. Coding allowed us to organize the data better and facilitate sensemaking (Yin, 2018).

We used the principles of MLOps identified in literature as an organizing framework: 1) continuous training, integration, and deployment, 2) continuous monitoring and feedback loops, and 3) transparency and reproducibility. These principles provided a natural fit since they stand at the core of MLOps, determine MLOps maturity, and have already served as a structure in our interviews. We then aggregated our findings for each ML project and assessed them using the previously identified levels of MLOps maturity (e.g., Google Cloud, 2020; John et al., 2021).

We then analyzed our data to potentially identify factors that would explain the different levels of MLOps maturity that we observed. For example, we looked at the challenges the ML projects faced and their unique context. As part of this process, we returned to the ML and MLOps literature to situate and challenge the emergent factors. Last, we handed our results to an ML engineer at AutoCorp, who evaluated the correctness and plausibility of our analysis from an insider's and practitioner's perspective.

5. Results

We assess the MLOps maturity in the three ML projects examined at AutoCorp in the following. We do so by individually assessing MLOps maturity along the areas of 1) continuous training, integration, and deployment, 2) continuous monitoring and feedback loops, and 3) traceability and reproducibility for each ML project. We further highlight salient challenges that complicated the application of more mature MLOps practices at AutoCorp. We conclude with an overall assessment of MLOps maturity across the ML projects. After that, we will discuss the factors that influence MLOps maturity. Table 4, Table 5, and Table 6 summarize our findings.

5.1 Continuous Training, Integration, and Deployment

Table 4. Summary of findings regarding continuous training, integration, and deployment

Case	Summary of Findings
ML-P1	**Moderate maturity:** Automated ML training; manual deployment
ML-P2	**Moderate maturity:** Automated ML training; manual deployment
ML-P3	**High maturity:** Fully automated ML pipeline (including deployment)
Cross-case	**MLOps challenges:** - Stability of the environment - Selection of appropriate MLOps tools - Integration tests - Verification of ML systems

A differentiated picture emerged when examining the principle of continuous training, integration, and deployment. Project ML-P3 has already built a **fully automated pipeline**, including training, integration, and deployment, which enables periodic updates of the ML system in an automated way. The projects ML-P1 and ML-P2 have automated large parts of the pipeline, including training and testing, but ultimately conduct **manual deployments** of the ML models. In ML-P1, the team currently pilots an automated ML pipeline (including training, integration, and deployment) which should simplify regular updates in the future. In ML-P2, a startup provides manual updates on-site and on demand.

A possible explanation could be that ML-P1 and ML-P2 are currently in a stabilization phase, where many updates to the ML model are required. During this phase, project teams typically discover **novel and changing parameters in the environment** that require adjustments to the ML model and potentially the pipeline (e.g., novel outliers or changing input data quality). Hence, a fully automated ML pipeline, including automated deployments, has thus far not been the focus. Another challenge during the early phases of the projects pertains to the **selection of MLOps tools** from a heterogeneous and dynamically evolving market, especially when the organization is still to develop valuable experiences and best practices. Furthermore, ML-P1 reported challenges regarding setting up **integration tests** for the ML system (e.g., testing the integration with camera hardware in the loop). In addition, ML-P3 explicitly highlighted the challenge of **verifying ML systems** for quality assurance in manufacturing according to industry standards. Current industry standards regarding quality assurance do not recommend

specific measures for ML systems. Hence, teams must invest much effort upfront to collect and develop verification criteria and high-quality test datasets. The lack of standards and best practices pose an additional barrier to automating testing for MLOps. ML projects can further attenuate this issue by deploying the ML system for augmentation rather than full automation. Hence, a human user can correct possible wrong outputs of the ML system.

5.2 Continuous Monitoring and Feedback Loops

Table 5. Summary of findings regarding continuous monitoring and feedback loops

Case	Summary of Findings
ML-P1	**Moderate maturity:** Manual monitoring of several measures
ML-P2	**Low maturity:** Manual monitoring through user feedback in-use
ML-P3	**Moderate maturity**: Manual monitoring of several measures
Cross-case	**MLOps challenges:** - Configuring sensitivity of monitoring - Appropriateness of external tools - Obtaining ground-truth feedback - Streamlining new data labeling

The principles of continuous monitoring and feedback loops are still at an early stage. All three projects regularly collect several measures for the performance of ML systems (e.g., logging the confidence of the ML system's predictions). However, ML engineers currently need to check these measures **manually periodically**. In this sense, **no "intelligent monitoring" takes place** yet, as described in two interviews. Such intelligent monitoring would include advanced measures such as data drift detection and automatically raising alarms based on critical deviations from predefined thresholds. However, the interviewees expressed uncertainty about configuring the **sensitivity of the monitoring** to detect relevant variations in time while not triggering a flood of false alarms. Another challenge we identified pertains to the **lack of appropriate MLOps tools on the market** that fit the **domain-specific requirements** for monitoring. Taking matters into their own hands, ML-P1 has recently pioneered an in-house solution to detect data drift in manufacturing image data.

Another critical aspect of monitoring is to receive **ground-truth feedback** on the actual real-world performance of the ML system to enable feedback loops (e.g., whether an error predicted by the ML system represented an error in the real world). In ML-P3, regular samples are taken and automatically assessed using accurate (yet resource-intensive) sensor measurements, which can accurately and timely reflect the **real-world performance of the ML system**. In ML-P1 and ML-P2, however, such a unique option is not available at hand. Instead, ML-P1 and ML-P2 receive feedback via established service desk channels where users can flag potential issues. Additionally, ML-P1 regularly collects and labels new data to provide a benchmark for real-world performance. ML-P1 selects this data based on insights from their monitoring (e.g., data drift detection). In the future, ML-P1 and ML-P2 could also explore the role of human workers in giving direct feedback to the ML system (e.g., pushing a button when the ML system produces a wrong output).

To conduct retraining based on new data, ML-P1 and ML-P2 needed to find solutions to **streamline the labeling of new data**, especially when labeling is non-trivial and requires domain knowledge. The main challenges for this are determining which new data should be labeled for the training and deciding to what extent the business should be involved. ML-P1 established standard processes to integrate the business to address this issue, resulting in a frictionless process to label new data for retraining. Moreover, to simplify the selection of new training data, ML-P1 indicated exploring the concept of active learning in the future, meaning the ML system proactively selects new training data according to its likelihood to maximize learning benefits. Similarly, the external developer team of ML-P2 visits the local site to choose and label new data. In contrast, ML-P3 can conveniently use the abovementioned, regularly taken samples as new and already labeled data.

5.3 Traceability and Reproducibility

Table 6. Summary of findings regarding traceability and reproducibility

Case	Summary of Findings
ML-P1	**High maturity:** Sophisticated versioning and meta-data tracking
ML-P2	**Moderate maturity:** Limited traceability due to external development
ML-P3	**High maturity:** Sophisticated versioning and meta-data tracking
Cross-case	**MLOps challenges:** - Common standards for MLOps tools - Leverage synergies (e.g., systematic sharing of ML artifacts)

A very advanced state emerged throughout the projects regarding the principle of traceability and

reproducibility of ML systems. Tools for **versioning ML-specific artifacts** and **experiment tracking** are applied in all three ML projects, ensuring the traceability and reproducibility of ML systems. This is likely because **many MLOps tools are available** on the market that supports these features today. Furthermore, in the context of automotive, creating the corresponding transparency here is necessary. However, as ML-P2 is **externally developed and operated**, this traceability and reproducibility are not seamlessly accessible to AutoCorp.

Another aspect concerns whether teams can use the provided traceability and reproducibility within projects to exchange ML models and data between projects systematically. Such exchanges require **standards regarding MLOps tools** and **processes to systematically exchange ML artifacts**, such as ML models, standard features, and training data. However, setting such standards requires delicately balancing the specific requirements of each project with the benefits of leveraging synergies. For computer vision projects, AutoCorp addresses this with a group-wide computer vision platform. This platform aims to achieve synergies within the firm by, among other things, sharing trained ML models.

5.4 Overall Assessment of MLOps Maturity

Based on our findings, the projects **ML-P1** and **ML-P2** feature a moderate MLOps maturity as they currently **automate ML training** and partially provide traceability and reproducibility. Nevertheless, both projects face a long way toward fully automated ML operations, including automated deployments and feedback loops. In contrast, **ML-P3** already features a higher MLOps maturity as it employs a **fully automated ML pipeline** and measures to collect feedback automatically in production. However, ML-P3 has not reached the highest maturity level as it is yet to employ advanced monitoring solutions, such as data drift detection. Apart from the positive experiences at ML-P3, we conclude that AutoCorp currently experiences a noticeable gap between ML development and operations. As a result, updates of ML systems involve a very high level of manual effort and face potential delays. In the following, we further discuss our findings from the three ML projects and derive factors that influence MLOps maturity.

6. Discussion

This present study aims to further shed light on the factors that influence the different levels of MLOps maturity observable in practice. To that end, we conducted a case study on MLOps maturity with three ML projects at AutoCorp. In line with John et al. (2021), our study observes different levels of MLOps maturity across the projects, ranging from laborious manual deployments to fully automated ML pipelines. Our analysis of the case study and literature suggests several **factors that likely facilitate or inhibit MLOps maturity** to explain the observable differences in MLOps maturity (cf. Figure 2). In the following, we discuss these factors alongside this study's implications for research and practice.

6.1 Factors Facilitating MLOps Maturity

We identified three factors likely to facilitate higher MLOps maturity. First, higher **organizational experience with MLOps** likely facilitates higher MLOps maturity. AutoCorp, like many other firms, initially has had little experience with MLOps. Together with a general lack of best practices regarding MLOps (e.g., Martínez-Fernández et al., 2022; Paleyes et al., 2022), AutoCorp lacked guidance for critical aspects of MLOps, such as defining quality gates for ML systems, or determining the update frequencies for ML systems. Consequently, all ML projects examined at AutoCorp decided to initially start with much manual work (low MLOps maturity) as they first wanted to gain the necessary experience. However, after some time, all ML projects eventually shifted toward higher MLOps maturity by piloting and implementing more automated processes (e.g., automatic data drift detection). This observation corresponds to prior findings suggesting the evolving nature of MLOps maturity in ML projects (Lwakatare et al., 2019; Mucha et al., 2022) and the importance of building experience and expertise for MLOps (Kreuzberger et al., 2023; Nahar et al., 2022).

Second, the **appropriateness of existing MLOps tools** likely facilitates higher MLOps maturity. For example, all ML projects examined reported no challenges concerning traceability and reproducibility, as those represent rather generic features that many commercially available MLOps tools support. In contrast, one ML project team screened several commercial MLOps tools for data drift detection but ultimately had to develop their solution. This finding aligns with recent observations that the market for MLOps tools is still emerging (AI Infrastructure Alliance, 2022). Another aspect to consider when judging the appropriateness of existing MLOps tools is their compatibility with the existing landscape (Granlund, Kopponen, et al., 2021; Ruf et al., 2021).

Third, high **business criticality of the use case** likely facilitates higher MLOps maturity. As Lwakatare et al. (2019) argue, more critical use cases that pose considerable business risk and do not allow

for human intervention require more mature MLOps practices, such as highly automated ML pipelines and advanced monitoring. This allows teams to detect performance issues early and update ML systems in a reliable and reproducible way. Similarly, at AutoCorp, we have observed a strong desire to employ mature MLOps practices as many of the use cases developed deal with critical tasks like quality assurance.

6.2 Factors Inhibiting MLOps Maturity

We further identified four factors likely to inhibit higher MLOps maturity. First, a high **inherent complexity of the underlying ML model** likely inhibits higher MLOps maturity. We suggest that the ML model's complexity comprises aspects such as model architecture and size, interpretability, and data requirements. For example, when working with unstructured data, ML-P1 and ML-P2 found automating tasks related to data preparation and annotations more challenging. In comparison, ML-P3, who worked with structured sensor data, faced significantly fewer challenges. Furthermore, ML models based on deep learning may further complicate the debugging and error tracing of the ML system due to their opaque nature (Martínez-Fernández et al., 2022). Thus, it may be significantly harder to automate and streamline the retraining and updating process.

Second, insufficient **quality of new data** likely inhibits higher MLOps maturity, as it complicates updating the ML model with new training data. This aligns with prior observations that pointed to data-related challenges in MLOps (Lwakatare et al., 2019; Ruf et al., 2021; Symeonidis et al., 2022). For example, ML-P1 and ML-P2 could not readily access new data for retraining. The data must first be manually screened, selected, prepared, and labeled, which impedes continuous retraining and increases the time for updates. ML-P1 coped with this issue by successfully establishing processes to engage business experts to streamline new data preparation systematically. Even though MLOps emphasizes the automation of workflows, this finding highlights the important role of engaging human users in the continuous development of ML systems (Grønsund & Aanestad, 2020; Waardenburg & Huysman, 2022).

Third, a high **environmental complexity** surrounding the ML system likely inhibits higher MLOps maturity, particularly in early project phases. For example, ML-P1 hesitated to deploy a fully automated ML pipeline before they fully stabilized the ML system in a complex and dynamic production environment. During initial deployments at different production lines, the team often recognized novel environmental parameters that required modifications to the ML model or its supporting systems. This confirms the challenge of setting up a stable, automated ML pipeline (Ruf et al., 2021). Those challenges might further intensify when teams deploy ML systems to multiple locations where they have to deal with different environments (Weber et al., 2022).

Fourth, the **idiosyncrasy of the use case** is likely to inhibit higher MLOps maturity. We argue that high idiosyncrasy decreases the likelihood of finding suitable off-the-shelf solutions, research, or best practices regarding MLOps. These forces project teams to custom-develop novel concepts and solutions. For example, one project at AutoCorp researched and developed its data drift detection due to a lack of practical commercial MLOps tools. Similarly, Ackermann et al. (2018) reported on an ML project in the public sector where they had to custom-develop due to the use case's idiosyncratic requirements. As another example, interviewees reported challenges of verifying the ML system in the context of quality assurance. In other contexts, such as banking, regulators have already provided clear guidelines for verification (Paleyes et al., 2022).

6.3 Implications for Research and Practice

We contribute to research on managing and organizing AI with factors that potentially explain the different maturity levels of MLOps observable in practice. We confirm prior findings highlighting an **organization's experience** in successfully deploying

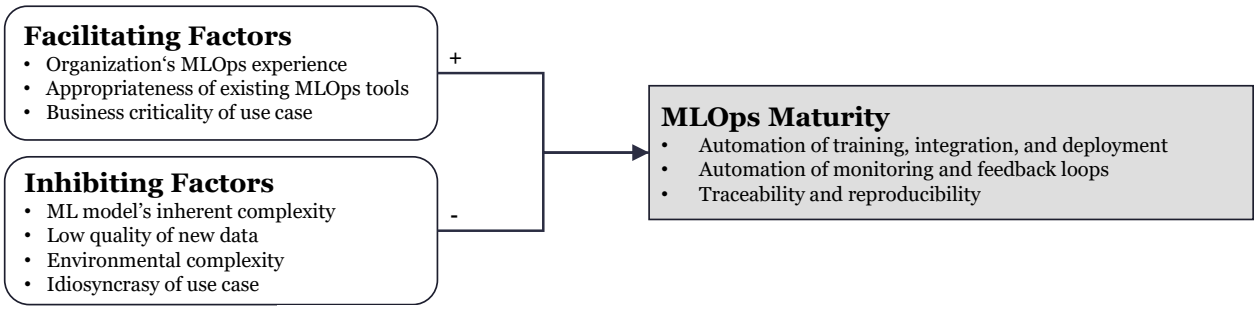

Figure 2. Factors influencing MLOps maturity

ML systems (e.g., Shollo et al., 2022). We further found that the emergent and heterogeneous **market of MLOps tools and services** (AI Infrastructure Alliance, 2022) likely plays an important role by providing (or not providing) appropriate MLOps tools to projects. In addition, our findings suggest projects should consider **a range of project-inherent factors** when employing MLOps workflows and principles, such as the ML model's complexity and the quality of new data. This observation further implies that specific ML systems (e.g., deep learning in an unstable environment) are inherently more challenging to operate, as they complicate using more mature MLOps practices. These project-inherent factors might also provide an **alternative explanatory pattern** to why some organizations find it hard to deploy specific ML systems (e.g., Benbya et al., 2020; Paleyes et al., 2022). We invite future research to create a deeper understanding of the different characteristics and contingencies of ML projects and the implications they pose for managing and organizing AI.

We also contribute to an emergent stream of empirical research on MLOps. As indicated in prior literature (Granlund, Stirbu, et al., 2021; John et al., 2021; Lwakatare et al., 2019), we could observe that **MLOps maturity typically grows over time** as teams gain experience and the ML workflow reaches sufficient stability throughout a project. In addition, even though MLOps emphasizes the automation of workflows (e.g., Kreuzberger et al., 2023), our findings highlight the **vital role of human users** (Waardenburg & Huysman, 2022) in the continuous development and operation of ML systems (e.g., to select and label new training data). Hence, despite increasing automation through MLOps, human users seemingly remain relevant in maintaining ML systems through new data preparation and providing feedback (Grønsund & Aanestad, 2020; Lyytinen et al., 2021).

Our findings may also serve as a valuable reference for practitioners striving to employ MLOps workflows and principles. First, we recommend practitioners consider their expertise, the appropriateness of existing MLOps tools, and project-inherent factors when employing MLOps in their projects. Second, considering the role of expertise, practitioners may start with lower maturity levels of MLOps before gradually attaining more advanced levels. For example, only after manually monitoring an ML system over several weeks might one consider automating the monitoring through automated alarms.

7. Conclusion

MLOps provides a set of workflows and principles that could help address the various and unique challenges in developing and operating AI systems based on ML. Our case study in an automotive firm suggests several factors influencing MLOps maturity, such as the ML model's complexity, the quality of new data, and the appropriateness of existing MLOps tools. We advance the current discourse on managing and organizing AI and help to explain the different adoption of MLOps in practice. We invite future research to further explore project-inherent factors as contingencies in ML projects, the role of human users in MLOps, and the process of how MLOps matures within projects and organizations.

References

Ackermann, K., Walsh, J., De Unánue, A., Naveed, H., Navarrete Rivera, A., Lee, S.-J., Bennett, J., Defoe, M., Cody, C., & Haynes, L. (2018). Deploying machine learning models for public policy: A framework. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK.

AI Infrastructure Alliance. (2022). *AI Infrastructure Ecosystem of 2022*. https://ai-infrastructure.org/ai-infrastructure-ecosystem-report-of-2022/

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial Intelligence in Organizations: Current State and Future Opportunities. *MIS Quarterly Executive, 19*(4), ix-xxi.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *Mis Quarterly, 45*(3), 1433-1450.

Böttcher, T. P., Weber, M., Weking, J., Hein, A., & Krcmar, H. (2022). Value drivers of artificial intelligence. 28th Americas Conference on Information Systems (AMCIS), Minneapolis, MN.

CD Foundation. (2022). *MLOps Roadmap 2022*. Retrieved January 11th from https://github.com/cdfoundation/sig-mlops/blob/main/roadmap/2022/MLOpsRoadmap2022.md

Dolata, M., Crowston, K., & Schwabe, G. (2022). Project Archetypes: A Blessing and a Curse for AI Development. 43rd International Conference on Information Systems (ICIS), Copenhagen, Denmark.

Fitzgerald, B., & Stol, K.-J. (2017). Continuous software engineering: A roadmap and agenda. *Journal of Systems and Software, 123*, 176-189.

Gall, M., & Pigni, F. (2022). Taking DevOps mainstream: a critical review and conceptual framework. *European Journal of Information Systems, 31*(5), 548-567.

Google Cloud. (2020). *MLOps: Continuous delivery and automation pipelines in machine learning*. Retrieved January 11th from https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Granlund, T., Kopponen, A., Stirbu, V., Myllyaho, L., & Mikkonen, T. (2021). Mlops challenges in multi-organization setup: Experiences from two real-world cases. 2021 IEEE/ACM 1st Workshop on AI

Engineering-Software Engineering for AI (WAIN), Madrid, Spain.

Granlund, T., Stirbu, V., & Mikkonen, T. (2021). Towards regulatory-compliant MLOps: oravizio's journey from a machine learning experiment to a deployed certified medical product. *SN computer Science, 2*(5), 1-14.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems, 29*(2), 101614.

Hemon-Hildgen, A., Rowe, F., & Monnier-Senicourt, L. (2020). Orchestrating automation and sharing in DevOps teams: a revelatory case of job satisfaction factors, risk and work conditions. *European Journal of Information Systems, 29*(5), 474-499.

Hill, C., Bellamy, R., Erickson, T., & Burnett, M. (2016). Trials and tribulations of developers of intelligent systems: A field study. 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), Cambridge, UK.

John, M. M., Olsson, H. H., & Bosch, J. (2021, 1-3 Sept. 2021). Towards MLOps: A Framework and Maturity Model. 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Palermo, Italy.

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access, 11*, 31866-31879.

Laato, S., Mäntymäki, M., Minkkinen, M., Birkstedt, T., Islam, A., & Dennehy, D. (2022). Integrating machine learning with software development lifecycles: Insights from experts. 30th European Conference on Information Systems (ECIS), Timișoara, Romania.

Leite, L., Rocha, C., Kon, F., Milojicic, D., & Meirelles, P. (2019). A survey of DevOps concepts and challenges. *ACM Computing Surveys (CSUR), 52*(6), 1-35.

Lwakatare, L. E., Raj, A., Bosch, J., Olsson, H. H., & Crnkovic, I. (2019). A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation. Agile Processes in Software Engineering and Extreme Programming, Montréal, Canada.

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems= humans+ machines that learn. *Journal of Information Technology, 36*(4), 1 –19.

Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A. M., & Wagner, S. (2022). Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM), 31*(2), 1-59.

Microsoft. (2023). *Machine Learning operations maturity model*. Retrieved January 11th from https://learn.microsoft.com/en-us/azure/architecture/example-scenario/mlops/mlops-maturity-model

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. SAGE.

Mucha, T. M., Ma, S., & Abhari, K. (2022). Beyond MLOps: The Lifecycle of Machine Learning-based Solutions. 28th Americas Conference on Information Systems (AMCIS), Minneapolis, MN.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization, 17*(1), 2-26.

Nahar, N., Zhou, S., Lewis, G., & Kästner, C. (2022). Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. Proceedings of the 44th International Conference on Software Engineering, Pittsburgh, PA.

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: a survey of case studies. *ACM Computing Surveys (CSUR)*, 1-26.

Riungu-Kalliosaari, L., Mäkinen, S., Lwakatare, L. E., Tiihonen, J., & Männistö, T. (2016). DevOps adoption benefits and challenges in practice: A case study. Product-Focused Software Process Improvement: 17th International Conference, PROFES 2016, Trondheim, Norway.

Ruf, P., Madan, M., Reich, C., & Ould-Abdeslam, D. (2021). Demystifying mlops and presenting a recipe for the selection of open-source tools. *Applied Sciences, 11*(19), 8861.

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Shollo, A., Hopf, K., Thiess, T., & Müller, O. (2022). Shifting ML value creation mechanisms: A process model of ML value creation. *The Journal of Strategic Information Systems, 31*(3), 101734.

Sjödin, D., Parida, V., Palmié, M., & Wincent, J. (2021). How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops. *Journal of Business Research, 134*, 574-587.

Symeonidis, G., Nerantzis, E., Kazakis, A., & Papakostas, G. A. (2022). MLOps-Definitions, Tools and Challenges. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV.

Testi, M., Ballabio, M., Frontoni, E., Iannello, G., Moccia, S., Soda, P., & Vessio, G. (2022). MLOps: A Taxonomy and a Methodology. *IEEE Access, 10*, 63606-63618.

Waardenburg, L., & Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization, 32*(4), 100432.

Weber, M., Pfeiler, M., Hein, A., Weking, J., & Krcmar, H. (2022). Deploying AI Applications to Multiple Environments: Coping with Environmental, Data, and Predictive Variety. 43rd International Conference on Information Systems (ICIS), Copenhagen, Denmark.

Wiedemann, A., Forsgren, N., Wiesche, M., Gewald, H., & Krcmar, H. (2019). Research for practice: the DevOps phenomenon. *Communications of the ACM, 62*(8), 44-49.

Wiedemann, A., Wiesche, M., Gewald, H., & Krcmar, H. (2020). Understanding how DevOps aligns development and operations: a tripartite model of intra-IT alignment. *European Journal of Information Systems, 29*(5), 458-473.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.