Universiteit Leiden

ICT in Business and
the Public Sector

Due Diligence for AI-based software systems

Lion Cassens
Student number: 2906953

1st supervisor: Prof.dr.ir. Joost Visser
2nd supervisor: Dr.-Ing Marc Hilbert

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

August 09, 2022

Abstract

Introduction: Artificial intelligence (AI) has an increasing impact on businesses and likely becomes more important to mergers and acquisitions. Nevertheless, little research has been done on the implications of AI-based software systems for Due Diligence. We examine how Due Diligence can identify AI within a software system, possible risks of AI-based software systems, and how Due Diligence can find such risks of a specific system.

Background: previous research around policy making, internal auditing, quality assurance, maturity, and software engineering provides insights into the implications of AI for the respective fields. While, to our knowledge, no research proposes a framework for Due Diligence on AI, we argue that insights from related fields can support the development of such a framework.

Method: we propose a three-part framework for Due Diligence on AI-based systems derived through the use of design science methodology. The first part of the framework consists of eight criteria to identify different AI techniques in a software system. The AI criteria are meant to form an entry point for further examinations with the framework. The second part is a benchmarking of machine learning software engineering best practices. We argue that a lack of best practices contributes to risks from an AI-based software system. The third part of the framework is a structured risk assessment with a catalogue of possible risks in nine areas. The goal of the structured risk assessment is to systematically identify the risks of an AI-based system within a merger or acquisition.

Results: the AI criteria and best practice benchmark were evaluated on a real-world Due Diligence case. Both framework parts were considered useful. However, some AI criteria and best practices were difficult to understand or answer. The structured risk assessment was evaluated with a case study on a fictional Due Diligence on an open source application for crowd detection. The participants indicated that the risk assessment increases their productivity, supports them in identifying risks, and that they would use the assessment for future Due Diligence cases.

Discussion and conclusion: we found that a clear distinction between the existence or absence of AI within a system is difficult. Instead, Due Diligence should focus on the different technologies from the broad spectrum of AI and their implications for the system. The proposed AI criteria can support consultants in this challenge. Risks from AI-based systems for mergers and acquisitions span a wide variety of areas. We conclude that tools like the proposed structured risk assessment can support consultants in the identification of risks. However, a catalogue of risks is likely never complete, and developments in the field and changing regulations require regular revisions of AI criteria, benchmarks, and risks.

Acknowledgements

I want to thank my supervisors, Prof. Joost Visser and Dr. Marc Hilbert, for their support and freedom throughout the research project. With great discussions and suggestions, both contributed to this master thesis.

Furthermore, I want to thank Dr. Erik Oltmans and his remarkable team for the opportunity to experience Due Diligence first-hand within the industry. The internship shaped not only my research project but was also a great personal experience for me.

Finally, I want to thank my family for all their love, my friend Werner for all the wisdom, and my friends from Oldenburg who made my whole student life much more exciting.

Contents

1 Introduction

With this document, we aim to better understand the implications of AI-based software systems for Due Diligence and how to deal with such implications. Therefore, we identify possible risks of AI-based software systems and propose a framework for the Due Diligence of AI-based software systems.

Mergers and acquisitions (M&As) within the technology industry are on the rise, driven by, among others, the desire to enter new markets and to obtain intellectual property [Haller, 2021]. While providing an excellent opportunity for growth, more than half of all M&As remain behind their financial or strategic objectives [Gomes et al., 2013]. Thus, it is important to evaluate the risks of M&As before a decision about a transaction is made. Due Diligence (DD) is the examination of a company to understand such risks and the value of a company. Due Diligence should cover the company's strengths and weaknesses over a wide range of areas in an objective and independent manner [Angwin, 2001].

1.1 Due Diligence in the Software Industry

One area of interest in Due Diligence is the information technology (IT) of a potential target. The examination of IT is often called Technology Due Diligence and focuses on IT issues, the compatibility with the IT of the acquiring firm, and the IT department of the target [Bhagwan et al., 2018]. Within the technology or software industry, a company's primary product or asset is often a software product. When such a software system is the primary driver of an acquisition, a value and risk assessment of the system becomes crucial. A high amount of technical debt in software might reduce its value [Groot et al., 2012] and potential security or privacy issues could lead to legal claims. Thus, the focus shifts from the overall IT landscape to a particular software system. In the following, we refer to Due Diligence of a particular software system as Software Due Diligence (SDD).

Artificial intelligence (AI) is becoming increasingly important for many businesses. A report by McKinsey shows an increase of companies that use AI for at least one function or business unit by almost 25% from 2018 to 2019 [Cam et al., 2019]. As acquisitions within the software industry can be a source of innovation [Schief et al., 2013] and a means to obtain an AI-based system, more acquisitions of AI-based businesses can be expected. AI-based software systems, especially systems that infer knowledge from data (i.e. machine learning), have different implications for software development and businesses than traditional software. Therefore, AI-based applications pose new challenges to Software Due Diligence. Nevertheless, little research has been done on the specific challenges for Software Due Diligence of AI-based software systems within such transactions. With this research, we want to make a step toward a better understanding of Due Diligence for AI-based systems.

1.2 Research Questions

To understand the risks and value of an AI-based system, one has to determine the degree to which a system actually reflects artificial intelligence. Furthermore, one must know the possible sources of risks of an AI-based system, and how to identify and ideally measure these risks. Therefore, we derived the following research questions (RQs):

- RQ 1: How can Due Diligence confirm AI within a software system?
- RQ 2: What are the possible risks of AI-based systems in the context of mergers and acquisitions?
- RQ 3: How can risks of a specific AI-based system be identified and measured during Software Due Diligence?

By pursuing an answer to the stated questions, we contribute to the understanding and improvement of the Due Diligence process for AI-based software systems, especially in the context of M&As.

1.3 Methodology

We followed the design science methodology to develop a guiding framework to conduct Due Diligence on AI-based systems [Wieringa, 2014]. We defined the primary design problem as follows: improve the Due Diligence on AI-based systems by designing a framework which guides the Due Diligence process consultants can apply in order to make risks from AI-based systems transparent.

The development of the framework was grounded in literature and semi-structured interviews. After an evaluation during the first design cycle, practical insights led to an extension of the model during a second design cycle. The framework was evaluated with two case studies, one on an actual Due Diligence case and one on a simulated Due Diligence case. Finally, insights from the evaluation and model development were linked to the research questions. We describe the stakeholders and contribution arguments for the framework in chapter 3.

1.4 Outline

After this introduction which motivated the undertaken research, we describe Due Diligence and give an overview of artificial intelligence in the background chapter. Furthermore, the chapter reviews relevant literature on AI from related fields like auditing, quality assurance, and software engineering. The chapter also discusses the applicability of methods from those related fields for Due Diligence. Chapter 3 describes the developed framework to conduct Due Diligence on AI-based systems. The chapter begins with an overview of the context and the stakeholders of the framework, which was developed within the context of a consulting organization. It then describes the three parts of the framework, namely eight criteria to identify different types of AI, a software engineering best practice benchmarking, and a structured risk assessment. Afterwards, we describe the evaluation of the developed framework with two case studies. Chapter 5 discusses the results in relation to the research questions and highlights the limitations of this research. Finally, chapter 6 concludes the research and proposes future work.

2 Background

2.1 Due Diligence

Within mergers & acquisitions, Due Diligence (DD) refers to the examination of the target to understand the values and risks involved in the transaction and covers a variety of areas [Angwin, 2001]. A literature review by Bhagwan et al. identifies the following areas of Due Diligence: financial, legal, tax, environmental, operational, market, human resources, cultural, strategic, marketing, intellectual property, technology, and R&D [Bhagwan et al., 2018].

We refer to the entity driving the Due Diligence as the client. On the other hand, the company examined in the Due Diligence is referred to as the target. The client is often either a company who can benefit from synergies through a merger or acquisition or an investment company.

2.1.1 Technology and Software Due Diligence

Within the academic literature, technology or IT Due Diligence usually refers to an examination of the whole IT landscape of a target [Bhagwan et al., 2018]. For software companies and digital businesses with custom software, an in-depth analysis of the digital or software products should be conducted instead. Several companies offer such Due Diligence services, often referred to as software, technical, or digital due diligence. Turuk and Morić Milovanović describe the focus of Digital Due Diligence as concerning the "various aspects of enterprise digitization with a special emphasis on their future sustainability" [Turuk and Milovanović, 2020]. Some providers market such services under the term of Technology Due Diligence or as part of IT Due Diligence. For disambiguation, we refer to the examination of the whole IT landscape as Technology Due Diligence and the examination of a digital business and customized software as Software Due Diligence. To date, academic research on Software Due Diligence is limited.

2.1.2 Software Due Diligence in practice

Many providers of Software Due Diligence analyze the technical debt, which typically involves a static code analysis to obtain objective metrics. A brief analysis of available offerings of six Due Diligence providers [1] shows that at least 5 offer an automatic analysis of technical debt and software quality. All perform an analysis of the strategy or product roadmap. At least four providers consider the infrastructure, security and team capabilities. At least three provide offerings for financial implications of a software product, and at least two Due Diligence providers include legal and intellectual property considerations in their Software Due Diligence offerings.

[1] Selected Due Diligence providers in alphabetical order: Cape of Good Code, Deloitte, EY, Implement Consulting Group, KPMG, SIG

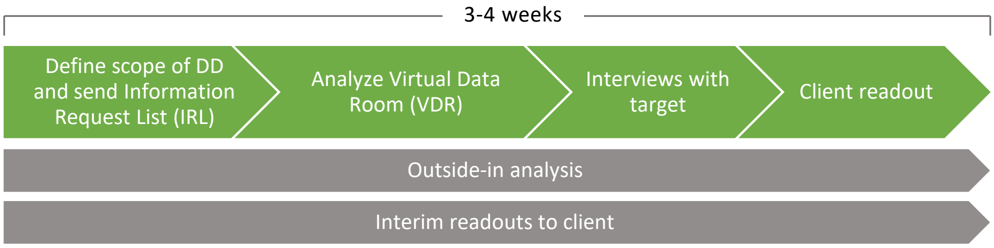

3-4 weeks
Define scope of DD and send Information Request List (IRL)
Analyze Virtual Data Room (VDR)
Interviews with target
Client readout
Outside-in analysis
Interim readouts to client

Figure 1: Due Diligence process as observed at a Due Diligence provider. Primary activities are highlighted in green and optional or supportive activities in grey.

.

Figure 1 provides a high-level overview of a typical Due Diligence process also used for technology and Software Due Diligence [2]. Tasks in green represent the primary process of Due Diligence. It is worth noting that the order of tasks can, vary and previous activities may be repeated at a later time due to new information. The process starts with the scoping of the DD. Based on the scope, required artefacts like documents, application access or source code are defined and requested through an Information Request List (IRL). Once the target provides access to the requested information through a Virtual Data Room (VDR), the DD provider analyzes the data. For Software Due Diligence, this step includes a static code analysis to identify the amount of technical debt, licensing issues, and potential security vulnerabilities. Typically, the analysis leads to open questions to be answered by the management or other appropriate roles during an interview with the target. Finally, the DD provider reports its findings to the client who commissioned the Due Diligence. Tasks in grey are optional and can accompany the whole process if necessary. The outside-in analysis may include interviews with industry experts or former employees to verify information and further insights [Rietveld, 2022].

2.2 Introduction to Artificial Intelligence

In order to understand the impact of AI-based systems on a transaction, a basic understanding of AI and its sub-fields is necessary. The term artificial intelligence has become more popular, with differing opinions on its meaning. Thinking of artificial intelligence, one could imagine a program able to win chess or Go against a human, a self-driving car, or a voice assistant within our smartphones. Expectations and underlying assumptions vary with the differing understanding of AI. Within the business world, one driver for the popularity of the term AI is the business value associated with its use. IBM found that AI is the second most critical technology regarding revenue impact since the COVID-19 pandemic [IBM, 2020]. Considering the value associated with AI, varying definitions of AI pose a challenge to the valuation and Due Diligence of companies that claim to facilitate AI. The following

[2] based on internal communication at EY-Parthenon, Amsterdam.

paragraphs give an overview of the field of AI to set the groundwork for further discussion on the value of AI and its impact on a strategic transaction.

2.2.1 Definition

Russel and Norvig describe two dimensions reflecting different efforts to define artificial intelligence. The first dimension differentiates between AI as a reflection of human behaviour versus rational behaviour. The second dimension differentiates between an internal or an eternal characterization of AI. An internal characterization defines intelligence through its "thought process and reasoning", while external characterization focuses on the behaviour exhibited by the system [Russell and Norvig, 2021].

An early and well-known approach to assessing whether or not a system constitutes AI was the Imitation Game, generally known as the Turing Test [Turing, 1950]. Turing proposes to consider a potentially intelligent machine as a black box. The test works as follows: a human posts written questions to the black box, receiving written answers. If the human can't differentiate the machines' answers from human answers, the machine passes the test [Turing, 1950]. Thus, the Turing Test considers artificial intelligence as a system that exhibits human behaviour. However, passing the Turing Test and the exhibition of human behaviour has not been the focus of most research [Russell and Norvig, 2021].

As described by Russel and Norvig, the dominant view on AI is that it should act rationally. This introduces the notion of a rational agent, always acting in order to achieve the best expected outcome. This approach generalizes well and benefits from good mathematical applicability, allowing the successful design of such agents [Russell and Norvig, 2021]. In order to strive for rationality, however, many algorithms are still inspired by the human learning process or by other intelligence behaviour exhibited in nature.

2.2.2 Origin and History

Russel and Norvig give an overview of the history of AI, summarized in the following. The first work representing AI is a model by Warren McCulloch and Walter Pitts from 1943 [McCulloch and Pitts, 1943]. Their model of artificial neurons was inspired by the human brain, where interconnected neurons respond to stimuli from neighbouring neurons, firing and thus, creating a stimulus themselves. Much of the subsequent research in AI focused on this neural network approach. First, algorithms were developed to learn with such artificial neural networks. Due to limited computational power and insufficient learning algorithms, the research around neural networks and other approaches to general-purpose AI slowed down, and research on expert systems increased. Expert systems leverage domain-specific knowledge. Many expert systems were rule or knowledge-based. Eventually, better algorithms, more available data, and increased computational power fostered the return of artificial neural networks and other machine learning methods. Such methods are still used today and allowed the breakthrough in the commercialization of AI [Russell and Norvig, 2021].

2.2.3 Sub-fields and applications

Contributions to artificial intelligence were inspired by several disciplines, including mathematics and especially statistics, psychology and neuroscience, linguistics, and computer science [Russell and Norvig, 2021]. Approaches to AI are equally widespread, resulting in several sub-fields. Here is an overview of important sub-fields:

- **Searching:** find a solution through simulated actions based on an abstract model (e.g. route-finding based on a graph representation of a country) [Russell and Norvig, 2021].

- **Knowledge representation:** deals with the representation of knowledge in ontologies, semantic nets, or rule systems [Russell and Norvig, 2021]. E. g. Wikidata, a structured knowledge base for Wikipedia.

- **Planning:** find a “sequence of actions to accomplish a goal”. Facilitates domain-independent algorithms and heuristics, enabled through a problem representation [Russell and Norvig, 2021].

- **Robotic:** field concerning with “physical agents that perform tasks by manipulating the physical world” [Russell and Norvig, 2021]. In a broader sense, this can include self-driving cars.

- **Speech Recognition and speech synthesis:** transformation of speech to text or another structured representation and the artificial synthesis of text to speech. Sometimes a prerequisite for Natural Language Processing and not always considered as AI on its own.

- **Natural Language Processing (NLP):** concerned with the automatic processing and representation of natural language for communication purposes and to learn from text [Russell and Norvig, 2021]. Recent language models pushed the boundaries of the field and its applications (e.g. BERT and GPT-3), benefiting machine translation, text summarization tasks, and sentiment analysis.

- **Computer Vision:** processing of visual input from cameras or comparable sensors to perceive the environment. Includes image classification and object recognition [Russell and Norvig, 2021]. Use cases include face detection and the recognition of speed signs and other objects in traffic.

- **Artificial General Intelligence:** tries to solve a variety of problems over different domains. Most research focuses on specific problems instead (narrow AI), as artificial general intelligence is considered challenging to achieve [Pennachin and Goertzel, 2007].

- **Machine Learning:** algorithms that learn from data or its environment [Russell and Norvig, 2021] and can be used for most sub-fields of AI and recently led to many remarkable achievements over many domains.

The overview shows how broad the field of AI is, with many fields overlapping or complementing fields. Especially Machine Learning (ML) supports many fields or problems within AI and has gotten much attention recently. Many breakthroughs in ML led to high-impact business models and products. The impact of ML on business highlights its importance for AI-based systems. Section 2.3 explains Machine Learning in more detail.

2.3 Machine Learning

Machine Learning uses algorithms to learn from training data. Such data can either be provided in the form of data sets or generated by exploration through an agent with its environment. In the former case, data can be annotated with the desired output the system should achieve. Such data is called labelled data, and the corresponding learning approach is called supervised learning. If data is provided but misses these annotations, unsupervised learning is applied. In the case of exploration, where data is not explicitly provided, we speak of reinforcement learning (i.e. the system learns from its interaction with the environment). The following sections further describe these types of Machine Learning.

2.3.1 Supervised Learning

In supervised learning, annotations or labels indicate the desired prediction that a model should make for every instance of the training data. Thus, the labels supervise the learning [Shalev-Shwartz and Ben-David, 2014]. The process of annotating data usually involves manual steps. Once a model is learned from the training data, previously unseen data (test data) is used to validate the accuracy of the model. If accuracy is solely measured on the training data, the performance might not reflect the actual performance of the model after its deployment. This is due to overfitting, where the model adopts the presented training data too strongly, losing the ability to generalize sufficiently for new, previously unknown data.

Typical supervised learning problems are classification and regression. A classification problem is where data instances should be assigned one of several predefined categories based on other instances for which the category is already known (labelled data). Examples of classification are spam detection

[Shalev-Shwartz and Ben-David, 2014] and classification of facial expressions [Happy and Routray, 2015]. In regression problems, a discrete or continuous outcome needs to be predicted for a certain instance based on other instances where such a discrete or continuous outcome is known (labelled data). Examples of a regression problem are the financial valuation of a building or the prediction of an income based on a person's educational background.

Common algorithms for supervised learning are decision tree algorithms, support vector machines, linear regressions, Bayesian networks, and artificial neural networks (to be discussed in section 2.3.4 on Deep Learning).

2.3.2 Unsupervised Learning

In unsupervised learning, a model is learned from unlabeled data. The desired outcome for a given data instance is not known in unsupervised learning. Therefore, no distinction between training and test data is made. An example of unsupervised learning is data clustering [Shalev-Shwartz and Ben-David, 2014]. Typical algorithms for unsupervised learning are principle component analysis (PCA) and k-means, which booth can be used to cluster data by its similarity [Ghahramani, 2004].

2.3.3 Reinforcement Learning

Reinforcement Learning tries to solve decision-sequence problems. The algorithm does not necessarily learn from data but from its interaction with the environment. An action generates a positive or negative reward, which allows the algorithm to learn the value of certain actions. The challenge with this is to associate future rewards with current actions, as actions are usually not independent from each other [Sutton and Barto, 2018].

Due to its sequential nature, Reinforcement Learning can be applied in, for example, games like chess and tic-tac-toe, and robotics [Sutton and Barto, 2018]. Recent breakthroughs in Reinforcement Learning involved Deep Learning, leading to the superior performance of AI over human champions in Go [Silver et al., 2016].

2.3.4 Deep Learning

Artificial neural networks are a machine learning approach where networks consisting of nodes in multiple layers connected with different weights are learned [Wang, 2003]. The first layer represents the input, while the last layer represents the output which in turn represents a prediction in classification or regression problems. During the training of the model, the weights are updated in order to minimize the error in the prediction or classification. Such a network can also be incorporated into reinforcement learning.

Once an artificial neural network has multiple layers besides the input layer, it is considered a deep neural network. One additional layer besides the input layer and the output layer technically fulfils the definition of deep learning. In practice, however, deep learning usually refers to dozens, hundreds, or thousands of such layers. Deep learning is very resource intensive, which is one of the reasons why only recently the field gained more momentum, even though the underlying techniques were long known [Russell and Norvig, 2021].

2.4 Artificial Intelligence in related fields

Even though we highlighted the importance of Due Diligence for AI-based systems in the introduction, little research has been done on Due Diligence for AI due to the novelty of widespread AI within businesses. This section looks at related fields and discusses the applicability of related approaches for Due Diligence.

2.4.1 Regulations

The European Commission appointed a high-level expert group on artificial intelligence (AI HLEG), which advises the EU on AI strategy [3]. Previous work of the AI HLEG focused on trustworthy AI, i.e., lawful, ethical and robust AI [European Commission, 2019a]. The expert group published ethics guidelines for such trustworthy AI [European Commission, 2019a], policy and investment recommendations [European Commission, 2019b], an assessment list for the trustworthiness of AI [Muller and Renda, 2020], as well as recommendations and sectoral considerations on the implementation of AI policies [European Commission, 2020].

The efforts toward stronger regulation of AI highlight an increasing need for auditing of AI, especially regarding the social impact of AI and its ethical aspects. To some extent, documents such as the assessment list for trustworthy AI can support internal auditing. The document acts as a checklist to ensure compliance with seven key requirements for trustworthy AI, earlier defined by the AI HLEG as follows [Muller and Renda, 2020]:

- **Human agency and oversight**: human agency means that the system should support the autonomy of the people affected by the system. Human oversight should be established to ensure human agency. Humans may be allowed to intervene or monitor the system.

- **Technical robustness and safety**: a system should be protected against hostile attacks, it should be accurate and results should be reproducible and reliable. Furthermore, fallback mechanisms should be established to minimize harm during a system failure.

[3] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

- **Privacy and data governance**: as already legally required by the general data protection regulation, AI systems must protect underlying data and the privacy of individuals affected by the system. Data quality and integrity must be ensured and documented.

- **Transparency**: requires traceability and explainability, which is hard to achieve with black-box methods like artificial neural networks. Furthermore, an AI system should never disguise itself, and the user should be made aware whenever it interacts with an AI system.

- **Diversity, non-discrimination and fairness**: AI systems should be checked for bias and allow access for all people without discrimination.

- **Societal and environmental well-being**: the social impact of an AI system should be monitored. All aspects of an AI system, including its development, should be environmentally friendly. The increasing computational power required for machine learning could pose a challenge to the sustainability of AI system developments. An AI system should not undermine democratic values.

- **Accountability**: an AI system should be auditable to create clear responsibility and accountability, and the impact of a system should be assessed.

Although the adherence to requirements for trustworthy AI likely becomes more and more important due to increasing regulations, such an assessment list has limited use for Due Diligence. The risks imposed by unethical, untrustworthy AI are only a subset of possible risks and implications that an AI-based system might have on a transaction. Financial, strategic, and technical aspects are not part of the assessment list, nor are they part of the work of the AI HLEG in general.

2.4.2 Internal Algorithmic Auditing for Accountability

Researchers from Google and the non-profit ”Partnership on AI” presented a framework for internal auditing during the development life-cycle of an AI-based system [Raji et al., 2020]. The framework's goal is to make an organization accountable for the AI systems they develop, and such accountability requires governance structures, which the framework puts in place. Auditing practices from regulated industries inspired the framework. They argue that traditional quality assurance alone is insufficient, as it does not check for ethical expectations. For example, quality metrics like loss or accuracy do not reveal information about potential social biases and the fairness of a system.

The framework consists of six stages which involve the auditors, the engineers of the audited system, and the responsible product teams [Raji et al., 2020]:

1. **Scoping**: In the first stage, intended use cases of the system are collected and subject to an ethical review based on company-wide AI Principles. Furthermore, the social impact of the system is accessed, focusing on possible unintended harm and mitigation strategies. The objective of the audit is defined based on the possible impact of the system.

2. **Mapping**: Stakeholders are identified, and their participation in the decision process is orchestrated (stakeholder buy-in). Interviews regarding the system development are held to identify gaps in the understanding of the systems impact and the decision-making process. The mapping stage follows an ethnographic methodology and is inspired by auditing processes of other industries. The start of the mapping stage should also trigger the start of a Failure Modes and Effects Analysis (FMEA), later used to prioritize risks in testing.

3. **Artifact Collection**: Required documentation for the development and testing of the application is collected based on a checklist of expected documents throughout the development cycle. Datasets should be described according to standardized data sheets. The model learned in the development process is described in a standardized model card. The model card includes underlying assumptions for the model development, expected behaviour, intended use case and risks.

4. **Testing**: During the testing stage, the auditors actively check the compliance with ethical values based on the risk prioritization from the FMEA. Testing can include adversarial testing, which tries to trick the system into unwanted behaviour. The tests should estimate the likelihood of risks to create an ethical risk analysis in combination with the severity of the risks.

5. **Reflection**: Test results from the previous stage are analyzed and checked against the company's AI principles. Risks are formalized and may be addressed in a mitigation plan. Due to the nature of the ethical nature of the audit, the risk analysis and mitigation plan focus on the social impact of the system.

6. **Post-Audit**: Subsequent to the audit, the reflection should inform a decision about the continuation of the system development and the implementation of the mitigation plan. Thus, the audit can potentially lead to the termination of a development project.

The framework gives companies who develop AI-based systems detailed steps to ensure accountability and thus, contributes to the compliance with prospective regulations regarding the use of AI, which could follow from the efforts of the high-level expert group on AI. However, the utility of this framework for due diligence is limited due to several factors: 1. the time horizon of due diligence is shorter than the time horizon of the framework, which accompanies the whole development life-cycle of an AI system. 2.

The framework is designed for internal use and access, while due diligence is conducted through an external party with potentially limited access to the target. Also, the internal approach may require more interaction with the development team than due diligence permits. 3. the framework focuses on ethical aspects and the social impact of AI, while due diligence should create a holistic view of all types of risks, including strategic and financial risks.

Nonetheless, the framework indicates several aspects to consider for AI systems that impact human lives. The system should be developed with its ethical implications in mind, and the social impact should be known. The suitable use cases of the model should be clear, and the limitations and implications of the dataset should be known. Model cards and data sheets could be created during the Due Diligence to make limitations and implications visible to the client.

2.4.3 Quality Assurance and Control for AI-based systems

The International Organization for Standardization (ISO) defines the quality of a software product by eight characteristics: functionality, efficiency, compatibility, usability, reliability, security, maintainability, and portability [ISO/IEC 25010, 2011]. Thus, software quality comprises not just the quality that a system exhibits towards its users (quality in use) but also internal quality characteristics, allowing the adoption of future requirements and continued satisfaction for its users. Software Quality Assurance takes systematic actions to ensure that software development and maintenance conforms to technical and managerial requirements [Schulmeyer, 2008, p. 27]. Quality control is part of quality assurance and evaluates the actual quality of the software product[Schulmeyer, 2008, p. 28 -29]. In the following, we look at quality assurance for AI-based systems and its similarities and differences to Due Diligence.

Felderer and Ramler [Felderer and Ramler, 2021] point out the need for a new approach to quality assurance for AI-based systems due to their data-intensive and self-adapting nature and their non-deterministic outcomes. They characterize AI-based systems on three dimensions:

- **Artifact Type**: An AI system consists of several artefacts which have to be considered by quality assurance for AI. Artefacts can be categorized as a system, framework, model, or data.

- **Process**: On the process dimensions, Felderer and Ramler state that an AI component can be either “developed in isolation or continuously by iteratively taking feedback from the deployed components into account” [Felderer and Ramler, 2021].

- **Quality Properties**: Felderer and Ramler consider data quality, quality in use, and software quality as relevant for every process and artefact type. Data quality is specifically important for machine learning, while software quality and quality in use are characteristics of every software system [Felderer and Ramler, 2021].

Due Diligence should verify that the quality of an AI-based system is sufficient and could therefore be inspired by approaches to quality control. Furthermore, different artefacts like the system, its model, and the data should be subject to Due Diligence for a holistic view of an AI-based system. Lastly, the Quality Assurance for the AI-based system itself could also be subject to Due Diligence to identify possible risks.

Batarseh et al. conducted a systematic literature review of quality assurance for AI and evaluated existing quality assurance methods on ten metrics [Batarseh et al., 2021]. They find that AI and AI quality assurance research is on the rise, with the latter currently focusing on explainable AI and trustworthiness. Further, they show that recent methods focus on specific sub-fields of AI within certain application domains [Batarseh et al., 2021]. The short period of Due Diligence makes it difficult to identify and adopt a domain and technology-specific framework. The only paper which describes a generic AI Quality Assurance method that scored positive on all ten metrics is discussed in the following.

Tao et al. describe several facets of Quality Assurance for AI [Tao et al., 2019]. They note that most AI systems are developed with science-based and not with engineering-based methods, leading to unclear quality requirements. Other challenges to quality assurance are limited data for validation and the uncertainty of outcomes. The paper also lists typical approaches to validate AI-based systems: classification-based testing can ensure diverse coverage of input data. Model-based testing for AI requires specific learning models. Other approaches are metamorphic testing, crowd-sourcing approaches, and rule=based testing. The authors acknowledge that the utilization of AI testing methods is still a challenge. Furthermore, they describe three dimensions of data validation for deep learning: 1. raw data quality checking, referring to data cleaning, quality monitoring, and evaluation of raw data. 2. validation processes for the generation of training data to improve its quality. 3. test data quality evaluation, which includes the coverage of test data [Tao et al., 2019]. Issues in these three aspects, raw data quality, training data quality, and test data quality, can create unidentified problems and create a risk for the future of an AI-based system. Therefore, all three aspects should be accounted for during Due Diligence on AI. Next to traditional software quality attributes, Tao et al. acknowledge specific quality attributes for AI-based systems. Depending on the software at stake, these can include correctness, accuracy and stability [Tao et al., 2019]. Due diligence should check if the target measures appropriate quality attributes.

2.4.4 Maturity Models for Artificial Intelligence

In recent years, efforts to develop models to evaluate the maturity of artificial intelligence in an organization have increased. Sadiq et al. [Sadiq et al., 2021] conducted a systematic literature review on AI maturity models. They found that such models can describe the current maturity level and highlight actions to improve or compare the maturity of different organizations. This is interesting for Due Diligence, as such models can potentially be used to identify weaknesses and validate the target's claims regarding their AI maturity. Sadiq et al. identified the following factors as critical for AI maturity: data, analytics, technology and tools, intelligent automation, governance, people, and organization. Unfortunately, most studies focus on a specific domain like logistics or drug research [Sadiq et al., 2021]. The literature review found that most maturity models focus on the maturity of Artificial Intelligence in an organization as a whole and not on an individual system. Furthermore, they conclude that the theoretical foundation of most models is insufficient and that more research on maturity models for AI is necessary [Sadiq et al., 2021].

Even though Due Diligence for AI-based systems should follow a holistic approach and incorporate organizational aspects, an AI maturity assessment that focuses on the organization differs in scope from such a Due Diligence. The early stage of research in maturity models and specific domain focus makes the use of such models for DD difficult. However, the critical factors identified for AI maturity might be a starting point to define subject areas for the Due Diligence of AI-based systems.

2.4.5 Best Practices for Machine Learning Software Engineering

Serban et al. identified 29 best practices for machine learning engineering and measured their adoption through a survey [Serban et al., 2020]. These practices consist of traditional software engineering practices, traditional practices modified for machine learning, and novel practices specifically for ML. They categorized each practice as data, training, deployment, coding, team, or governance-related. Furthermore, correlations between certain practices and effects (agility, software quality, team effectiveness, traceability) were identified. Later, Serban et al. introduced 14 new practices for trustworthy AI development based on the guidelines from the AI HLEG [Serban et al., 2021].

The survey questions and results for the best practices are publicly available. Furthermore, the authors provide the intent, motivation, applicability, description, and adoption for each best practice [4]. Overall, the best practices are actionable within AI development and possibly useful for an assessment or due diligence on an AI-based system and its environment. The available survey data allows a benchmark against the industry. However, as the best practices are limited to software

[4] `https://se-ml.github.io/practices/`

engineering for AI, strategic and financial aspects are missing. Therefore, a check against these best practices alone might not be sufficient for the Due Diligence of an AI-based system.

2.4.6 Conclusion

All considered articles from the related fields were published in or after 2019. Due Diligence, regulatory work, quality assurance, maturity models, and engineering practices regarding AI are all still in an early stage. None of the related work proposes a framework or methodology that can be used for Due Diligence without modifications or extensions as they are not actionable or not directly applicable for Due Diligence. Nonetheless, insights from related fields can be utilized to develop a Due Diligence framework for AI-based systems and knowledge about AI maturity models, best practices, and quality assurance can support a consultant in examining an AI-based system.

3 Framework for DD of AI-based systems

This section describes the framework designed to improve Due Diligence on AI-based systems during the DD process, intended to be used by consultants to make risks from AI-based systems transparent. The final framework was developed over two design iterations.

Based on the insights gained from the literature review and unstructured interviews with Due Diligence experts, we developed the first version of the framework for Due Diligence on AI-based systems. This first artefact included AI criteria to support the identification of AI components and their nature, as well as a benchmark of software engineering best practices. After a first evaluation during a real-world Due Diligence case, the framework was extended with a structured risk assessment. In the following, we define the context of this framework, describe its three parts (AI criteria, benchmark, and risk assessment), and give contribution arguments for each part of the framework.

3.1 Social context and stakeholders

The social context for the developed framework is the strategy consulting industry, which typically conducts Due Diligence for M&As. Stakeholders in this particular context are EY-Parthenon (sponsor of the research [5]), strategy consultants conducting Due Diligence, and companies targeted for a merger or acquisition (targets). EY-Parthenon is the strategy consulting division of EY, one of the big four accounting firms next to Deloitte, KPMG, and PwC. The research was conducted in the context of EY-Parthenon's Software Due Diligence services.

The concrete implementation of the framework within the evaluation followed the structure of typical assessment tools at EY-Parthenon, but can easily be transferred to other formats and universally used without company-specific knowledge.

3.2 AI criteria

As AI can be broadly defined and comprises several techniques (see chapter 2.2), a logical first step within the Due Diligence of AI-based applications is to actually check whether or not a system exhibits AI and to which category the approach used by the system belongs. Interviews with due diligence consultants underpin the demand for a supportive tool to identify AI: several consultants mentioned the increased use of the term 'AI' in marketing, as the use of AI is implicitly connected to higher revenue expectations and financial valuations. In order to support consultants in assessing claims of AI, a catalogue of 8 AI criteria was developed, each indicating a specific aspect or type of artificial intelligence (e.g. rule-based systems or machine learning).

[5] Sponsors within the design science methodology provide the resources for the research and set goals for the design goal [Wieringa, 2014]

Certain criteria have a relationship to each other. As an example, the use of supervised learning implies the use of machine learning, as supervised learning is a subset of the latter. Such relationships were incorporated in the interactive excel spreadsheet to give the user real-time feedback on its inputs, allowing the identification of conflicting statements and their correction. Table 1 gives an overview of the criteria and their relationships.

	Criteria	**requires**
1	rule-based system	none
2	knowledge-bases	none
3	search or optimization	none
4	machine learning	5 or 6 or 7
5	supervised machine learning	4
6	unsupervised learning	4
7	reinforced learning	4
8	deep learning	4

Table 1: List of criteria implying different AI techniques and their relationship to each other within the framework.

The user indicates for each criteria whether or not it applies to the system (i.e. if it is true). Furthermore, each criteria selected as true for the examined system contributes a short descriptive sentence to a summary of this first assessment part. It is important to note that we do not define AI as requiring to fulfill every criteria, but rather that each criteria indicates a certain type and extent of AI. Each criteria is described in more detail below with the contents provided to the user: the criteria itself, additional information provided to the user, the indication shown if the user decides that the criteria applies to the system, and a short text for the summary. Afterwards, we provide additional information for each criteria, which we do not provide to the framework user.

3.2.1 Criteria 1: rule-based system

Definition: the system makes decisions based on manually programmed rules.

Additional information: Example 1: A chatbot diagnosing an illness based on multiple choice questions, where certain combinations of answers lead to a diagnose or recommended actions based on simple rules defined by experts. Example 2: A simple point of speech (POS) tagger that identifies POSs based on a list of known verbs, nouns and the like.

Indication if applies: rule-based system

Contribution to summary: rule-based AI components usually offer very limited added value and cannot be seen as true AI unless very sophisticated, including hundreds of rules.

Rule-based systems are a means to incorporate expert knowledge into an automated system to solve a problem [Hayes-Roth, 1985]. Rule-based systems were considered AI already back in 1985 [Hayes-Roth, 1985]; however: rule-based systems are not necessarily complicated systems but rather require additional manual labour to solve additional problems. Due to the current hype around AI, rule-based systems which may not have been labelled as AI in the past may be marketed as AI now to gain more attention or increase the perceived value of a system. We do not consider rules derived from machine learning as rule-based systems within this framework, as such a learning process has different implications than human-defined rules.

3.2.2 Criteria 2: knowledge-bases

Definition: the system represents knowledge in an explicitly defined structure, allowing the system to draw conclusions [Russell and Norvig, 2021].

Additional information: often, such a system is tailored to a specific domain to capture expert knowledge, e.g., medical terms and their relationships. An example of a general knowledge representation is WikiData: explicitly structured information from the Wikimedia Foundation (responsible for Wikipedia). Another example is WordNet, which represents relationships between words, such as synonyms.

Indication if applies: knowledge-base

Contribution to summary: whether or not a knowledge-based component can be seen as sophisticated AI depends on its volume and ability to infer new information.

The category of systems utilizing a knowledge base is broader than rule-based systems. A knowledge base can be derived either manually or through the support of machine learning. We see the knowledge base independently from the process of its creation (e.g. manual work or machine learning). A transaction could include a knowledge base but not the technology the knowledge base was crafted with.

3.2.3 Criteria 3: traditional search or optimization algorithms

Definition: the system uses an algorithm explicitly written to solve a specific problem that does not require learning [Russell and Norvig, 2021].

Additional information: algorithms written for an explicit problem can be found in a wide variety of applications: this could be a constraint satisfaction problem (e.g., optimizing production plans to maximize profits while meeting certain requirements). Another example is the planning of an optimal path for a laser to create circuit boards without overheating (temperature constraint). Algorithms that search for a solution can also be route planning algorithms.

Indication if applies: search or optimization

Contribution to summary: traditional algorithms that solve search or optimization problems range from very naive approaches to more advanced, specialized algorithms.

The implications of search and optimization algorithms can vary broadly. This category includes publicly available algorithms like A* to find the shortest path to highly customized proprietary algorithms. Thus, the implications of such algorithms for a transaction vary as well. It is difficult to conclude the extent of intelligence and impact on the product value without examining the algorithm at hand.

3.2.4 Criteria 4: machine learning

Definition: the system uses a model that is learned from training data or draws conclusions from data in another, automated manner. Thus, the system learns from data [Russell and Norvig, 2021].

Additional information: such systems often consist of more generic algorithms that are applied directly to data to generate insights, that learn a model based on data, or process data otherwise to learn a model that is then used to work with new, unknown data. Example 1: Prediction of income for an individual based on other people's salary information. Example 2: A webshop proposes interesting items based on previously seen items or similar product properties.

Indication if applies: machine learning

Contribution to summary (if criteria 8: deep learning does not apply): the system uses traditional approaches to machine learning. This does not automatically mean the system is not innovative or exhibits no competitive advantage, but it does not use deep learning, which was responsible for many recent breakthroughs.

Machine learning can range from simple models learned with standard algorithms on free data sets to highly customized or fine-tuned algorithms on large proprietary data. Even a system that does not apply deep learning can be sufficient at a task and add value to a system. The following three criteria help to understand if the learning of a model is conducted under supervision, unsupervised, or through reinforcement. The final criteria checks for deep learning, which can be used in all three types of machine learning.

3.2.5 Criteria 5: supervised learning

Definition: training data is labelled: the desired outcome for a given training instance is known; thus, the learning process is supervised [Russell and Norvig, 2021].

Additional information: data labelling usually requires human effort but allows the use of many algorithms, such as regression analysis, support vector machines, decision tree algorithms, and neural networks. Examples of a system facilitated by supervised learning are image recognition with a model trained on a large corpus of images and the category they belong to, the prediction of income based on job or age, and spam detection models trained on known spam and legitimate emails.

Indication if applies: supervised learning

Contribution to summary: criteria does not explicitly contribute to the summary, but requires *criteria 4: machine learning* and thus, triggers its description for the summary.

3.2.6 Criteria 6: unsupervised learning

Definition: training data is not labelled: the desired outcome for a given training instance is unknown. Thus, the learning process is unsupervised [Russell and Norvig, 2021].

Additional information: Unsupervised learning tasks often involve the clustering of data. Example 1: Finding similar items in a webshop. Example 2: predict missing words based on a text corpus and word frequencies in relation to surrounding words. Typical algorithms are k-NN, minhashing, principle component analysis (PCA), single value decomposition (SVD), or artificial neural networks.

Indication if applies: unsupervised learning

Contribution to summary: criteria does not explicitly contribute to the summary but requires *criteria 4: machine learning* and thus, triggers its description for the summary.

3.2.7 Criteria 7: reinforced learning

Definition: the system solves a sequential decision problem and learns through feedback (rewards) given for a series of actions proposed by the system [Russell and Norvig, 2021].

Additional information: often, such problems relate to games or robotics, as both domains consist of subsequent actions, which influence the environment. Examples of such problems are autonomous driving or a chess computer that improves through playing chess, learning from losing, winning, and the points obtained in each game. Such a system needs the ability to trace back rewards to subsequent actions.

Indication if applies: reinforced learning

Contribution to summary: criteria does not explicitly contribute to the summary but requires *criteria 4: machine learning* and thus, triggers its description for the summary.

3.2.8 Criteria 8: deep learning

Definition: the system uses multilayered artificial (deep) neural networks [Russell and Norvig, 2021].

Additional information: Typically, high computation efforts are involved in the development. Such systems often require a substantial amount of training data. Such deep neural networks are always used in the context of either supervised, unsupervised, or reinforcement learning. A neural network is usually considered deep once it has more than one hidden layer. Most practical use cases of neural networks use deep neural networks. Deep Learning is often applied in high-dimensional problem spaces.

Indication if applies: deep learning

Contribution to summary: deep learning methods are responsible for many recent breakthroughs in AI research and practical applications. This indicates true AI and a potential competitive advantage. However, it should be checked if traditional algorithmic approaches could solve the underlying problem as well. Deep learning introduces many variables and uncertainties. It should only be applied if the problem is too complex for traditional algorithms.

3.2.9 Contribution arguments

The goal of Due Diligence is to understand the values and risks of a transaction [Angwin, 2001] 2.1. The AI criteria presented above can form an entry point for further research about the specific AI components identified. This is important to understand values and risks as the implications of different AI criteria differ. Furthermore, the second and third part of the framework were developed specifically for machine learning. These criteria prevent inappropriate use of the other parts as they support the consultant in distinguishing machine learning from other AI techniques.

3.3 Best practice benchmarking

The second part of this framework is based on the work by Serban et al. on software engineering best practices for machine learning, as described in chapter 2.4.5 [Serban et al., 2020] [Serban et al., 2021]. We argue that a lack of best practices increases unforeseen risks as this can have a negative impact on the agility, software quality, team effectiveness, and traceability of a machine learning system [Serban et al., 2020]. Serban et al. surveyed machine learning practitioners to which extent they adhere to these best practices, which allows to benchmark the degree a target follows best practices against Serban's sample.

3.3.1 Benchmark design

We benchmark the target on all six categories and the four effects identified by Serban et al. See table 2 for an overview of categories and effects.

Categories:	Data, Training, Coding, Deployment, Team, Governance
Effects:	Agility, Software Quality, Team Effectiveness, Traceablility

Table 2: categories and effects of best practices in software engineering for machine learning [Serban et al., 2021]

The original survey by Serban et al. targeted machine learning practitioners and their development teams, whereas our benchmark is designed for usage through an external party. Therefore, the wording of the survey question for each best practice was altered to mirror this external perspective on a target. For example, the statement [Serban et al., 2020]:

> “Our training objective is captured in a metric that is easy to measure and understand.”

was altered to:

> “Target captures the training objective in a metric that is easy to measure and understand.”

Therefore, the statements are clear to the user while still reflecting the original meaning. This should improve the correlation between self-evaluations of engineers during the survey and external evaluations from consultants during due diligence. Such correlation is required in order to use the survey data for a best practice benchmark during due diligence.

In order to compare the scores given during the Due Diligence with the ones sampled by Serban's survey, the type and scale of survey questions should be compatible. To keep the comparison transparent and simple, we used the same 4-point Likert scale as the original survey. Thus, a target can adhere to a best practice: ”not at all”, ”partially”, ”mostly”, ”completely”, or a practice can be answered as ”does not apply”. For some of the best practices, a short indication, when the practice might not apply is also provided.

3.3.2 Calculation of benchmark results

All benchmark results consist of a score given by the user and a comparison to the benchmark data. The results aggregate over the categories of best practices, effects, and requirements for trustworthiness. User scores are derived by converting the selected entries of the 4-point Likert scale into numerical values (i.e. “not at all” is converted to

1 and “completely” to 4). Practices that do not apply to the target are ignored. The survey results by Serban et al. serve as our benchmark data and need to be converted into a numerical format as well. We consider the median response of each best practice as the benchmark score for that practice. To compare the user scores against the benchmark values, we considered two different approaches:

The first approach averages the user and benchmark score for each aggregate (i.e. categories, effects, requirements for trustworthiness). The average user score for each aggregate is then compared with the benchmark average. E.g. the average user score for the category data will be compared to the average benchmark score for data. It is important to note that missing user scores can bias the results if their respective benchmark scores are all particularly high or low. For missing scores with high corresponding benchmark scores, the results are biased towards a negative benchmark, whilst missing scores for low benchmark scores result in inflated benchmark results. Thus, we advice to score the target on all practices when using this approach.

The second approach summarises the differences between the user and benchmark score for every best practice and divides it by the number of scores. This procedure is applied to every aggregate (e.g. the scores within the data category). A resulting aggregate score of 0 indicates average compliance with best practices, while a negative score indicates lower compliance and a positive score higher compliance respectively. This approach can deal better with situations where scores are missing, as those can easily be ignored in the procedure. However, we find this approach less intuitive and the results more difficult to convey to a Due Diligence client. In the first approach, the user score aggregates can also be viewed independently, while the second approach always puts them in context with the benchmark data.

3.3.3 Contribution arguments

The best practice benchmark allows a quantification of the engineering practices at the target and potential impacts from the practices. Such a scoring model is a common tool in Due Diligence to establish more objectiveness in the evaluation of a software product [Rietveld, 2022]. The benchmark gives the client an impression of the engineering capabilities regarding machine learning and highlights potential areas of improvement. An online catalogue of the best practices can be used as a starting point for improvements in the development process as it includes a more detailed description of each practice [6].

However, we acknowledge that a quantified scoring of best practices is not sufficient to understand the implications of an AI-based system for a transaction. Thus, we added a structured risk assessment to this framework after the first design iteration.

[6] Catalogue of software engineering for machine learning best practices: https://se-ml.github.io/practices/ [Serban et al., 2020]

3.4 Structured risk assessment

The third part of the framework is a structured risk assessment designed to guide the thought process over nine risk areas. The areas are based on research in related fields (see chapter 2.4), interviews with consultants, and the experiences gained from the validation step of the first design cycle (AI criteria and best practice benchmark). The first validation took place in the form of a case study on a real-world due diligence case. It became apparent that reporting important risks adds significant value to a purely scoring-based report. Furthermore, such a structured risk assessment can capture areas outside of the software engineering scope.

Whether or not a characteristic of a system constitutes a risk within a transaction depends on the investment thesis (i.e. the motivation or goal for a transaction). Suppose a target is acquired to obtain access to its data sets. In that case, the potential weaknesses of a machine learning algorithm might be a relevant risk for the system but a less relevant risk in the context of the transaction. On the other hand, if the goal of a transaction is to achieve synergy effects from using a model for an extended pool of users, then risks of insufficient generalization may occur even though the model works sufficiently on its current user pool. Therefore, we argue that the consideration of the investment thesis is paramount to a correct assessment of AI-based risks during a merger or acquisition.

Figure 2 shows a high-level overview of the risk areas for the structured risk assessment. As mentioned before, the investment thesis influences the severity of possible risks. Furthermore, insights from the best practice benchmark can be incorporated into the consideration of each possible risk area. For instance, a target with below-average data practices may be more likely to exhibit data-related risks. However, considerations should not be limited to the insights from the benchmark as the risk assessment has a broader scope. For each area, the framework provides potential concrete risks within the assessment spreadsheet. The user can indicate if such a concrete risk does or does not apply or if it partly applies. A second field is designated for notes regarding the consideration of the user for the concrete risk. Each AI-related risk area also includes a note field for any other risk which was not explicitly described before. This stimulates the user to not limit its considerations to the risks mentioned in the assessment, as such a risk catalogue can never be completely exhaustive.

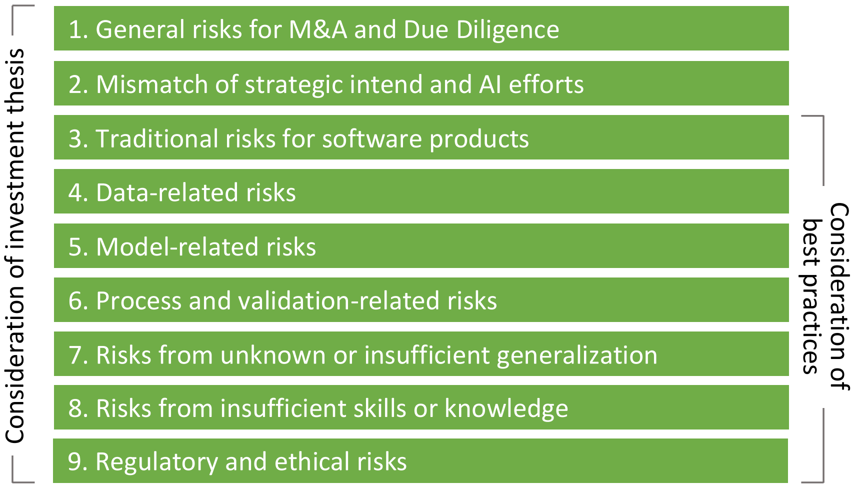

Consideration of investment thesis
1. General risks for M&A and Due Diligence
2. Mismatch of strategic intend and AI efforts
3. Traditional risks for software products
4. Data-related risks
5. Model-related risks
6. Process and validation-related risks
7. Risks from unknown or insufficient generalization
8. Risks from insufficient skills or knowledge
9. Regulatory and ethical risks
Consideration of best practices

Figure 2: overview of potential risk sources for AI-based systems and important considerations for the evaluation of potential risks within mergers and acquisitions.

In the following, we describe each potential risk area and the possible concrete risks of those areas. For the application of the framework, we omitted the description of each concrete risk after the first paragraph to ensure a balance of provided information and simplicity of the framework. Therefore, the first paragraph of each risk contains the general description, while the subsequent paragraph gives background information, embed scientific literature, and reasoning for including the risk.

3.4.1 Area 1: general risks for M&As and Due Diligence

This area describes the general risks of mergers and acquisitions and Due Diligence. As these risks do not focus explicitly on risks of AI-based systems, we did not allow notes or other input from the user. The area is listed for completeness and to prime the user for a potential influence of such general risks on the remaining categories.

We identified a negative business outcome of a merger or acquisition as a possible risk in this area as many mergers and acquisitions stay behind their financial objectives [Gomes et al., 2013].

Secondly, we acknowledge the risk of incorrect information or a wrong valuation basis: provided information may be incorrect, either due to errors in data or in an attempt to hide weak elements of a business. Therefore, Due Diligence should validate all information and ensure that a transaction does not rely on wrong assumptions. The user should keep this in mind during the risk assessment.

3.4.2 Area 2: mismatch of strategic intent and AI efforts

These risks relate to missing alignments of AI efforts with the overall product strategy and other strategic risks. We argue that this area is especially important as an extension to the previously described best practice benchmark, as the strategic view exceeds the scope of the best practices, which are grounded in engineering rather than strategic aspects of machine learning. Furthermore, potential strategic risks are not exclusive to machine learning but largely to AI in general.

a. AI efforts don't support the product strategy: AI becomes more and more important for software applications. However, this can lead to AI developments for the sake of development without a clear strategic intent.

Especially exploratory R&D activities with AI could be prone to fail in value creation. Not every AI project reaches the production level. Westenberger et al. identified common issues that contribute to the failure of such projects: unrealistic expectations, use-case related issues, organizational constraints, lack of critical resources, and technological issues [Westenberger et al., 2022]. During Due Diligence, it is important to consider the actual value AI adds to a system.

b. Weak competitive advantage: Even if AI efforts support the product strategy, its competitive advantage could be weak due to easy reproducibility. If competitors have easy access to training data and comparable algorithms, the competitive advantage might be weak.

More and more algorithms and libraries are becoming readily available as off-the-shelf commodities. Whole models are accessible through platforms like huggingface.co. If both the algorithm and the dataset are freely available or easily replaceable, then competitors can offer a similar value proposition.

c. Other AI risks regarding the product strategy: this section does not contain a description in the framework, as it allows the user to add any other considerations to the area that are not captured by one of the previous risks. The same applies to the last section of all following risk areas.

3.4.3 Area 3: traditional risks for software products

Risks for traditional software are still relevant for AI-based systems. Within this risk assessment, such risks should only be considered regarding the AI components of the system.

a. Risks from technical debt: technical debt in AI components may be higher than in regular software components due to the recent rise and novelty of widespread AI technology.

Technical debt in machine learning can be seen as a broader concept than technical debt in traditional software due to its data-driven nature. See Sculley et al. for an overview of machine learning-specific technical debt [Sculley et al., 2015]. However, as this area covers traditional risks for software development, we focus on technical debt as generally understood on dimensions like debt from low code quality, architecture, and testing (e.g. low test coverage) [Tom et al., 2013].

Lenarduzzi et al. argue that software quality in AI applications is comparably low and points out multiple reasons for such quality issues [Lenarduzzi et al., 2021]. Impressions of a few AI-based systems examined by consultants during this research strengthen the claim of lower code quality in AI components. Therefore, we find it important to specifically check the technical debt of AI components in a system during Due Diligence.

b. License or intellectual property issues: whenever external code is used, it is paramount that the required rights are obtained. For AI-based applications, this also includes the permission to use data for model training.

Even though a legal audit for a transaction is not the focus of Due Diligence, potential licensing or intellectually property issues should be briefly examined due to the data-driven nature of many AI applications (i.e. machine learning). The company Clearview AI, for example, developed facial recognition software and was fined by the United Kingdom and Italy for the use of images from social networks without the consent and knowledge of the depicted people [McCallum, 2022].

c. Cyber security vulnerabilities: security vulnerabilities can lead to data breaches or malicious behaviour. For machine learning, malicious input data may be used to trigger unwanted behaviour.

Biggio et al. showed how machine learning models could be attacked through slightly manipulated inputs. Such an attack can cause a system to classify objects falsely, for example, to evade spam or malware detection [Biggio et al., 2013]. The potential impact of such risks depends on the vulnerability of the AI-based systems to manipulation and the use case (e.g. potentially lower risk in entertainment software compared to applications in the health sector).

3.4.4 Area 4: data-related risks

Machine learning relies heavily on data, and the performance depends partly on data quality, and this shifts the focus from the source code to the data.

a. Insufficient raw data quality: raw data should be of high quality or cleaned before use. If raw data is biased or contains too much randomness, the model might not perform well. Furthermore, raw data must be available in sufficient quantity.

Tao lists "raw data quality checking" as one of three dimensions of data quality assurance in deep learning [Tao et al., 2019]. We argue that Due Diligence should check possible risks from insufficient raw data quality.

b. Test data does not reflect real-world conditions: Performance on test data may be significantly higher than real-world performance if test data does not reflect real-world conditions of the use case.

AI-based systems should be tested in the environment they are supposed to operate [Tao et al., 2019]. High accuracy on test data could be deceiving and ground for wrong assumptions within mergers and acquisitions. Therefore it is important to validate if evaluations of a system on test data are likely to reflect real-world performance, especially if an application is not yet used in production.

c. Uncertain availability of updated data: real-world conditions might change (concept drift). Therefore, a machine learning model needs to be updated regularly, which requires up-to-date training data in the future.

Context drift is a situation where the underlying assumptions or concepts of a problem change over time [Schlimmer and Granger, 1986]. An example of concept drift is spam detection in emails, as the phrasing of spam emails is likely to change over time. Depending on the investment thesis, the ability to adapt to such concept drifts can be more or less relevant. If the learning algorithm and not the trained model is of interest to the transaction, then updated data might be less relevant than if the value of the transaction lies in an up-to-date model in a use case with concept drift.

d. Other data-related risks: no description provided. Within the framework, this section is meant to stimulate the user to think about other risks related to data.

3.4.5 Area 5: model-related risks

Model-related risks refer to risks from the chosen algorithm and the learned model itself.

a. Inappropriate model choice: a machine learning model should not be unnecessarily complex as complexity increases the risk of unexpected behaviour. On the other hand, a model should achieve sufficient performance for the use case.

Checking for the risk of an inappropriate model choice is a generalization of the best practice to "employ interpretable models when possible" [Serban et al., 2020]. We argue that a less complex model simplifies maintenance and reduces the risk of unrecognised bias. When a use case requires performance only achievable with more complex models a naive model may be an appropriate choice. Therefore, the model appropriateness depends on the current use case and the investment thesis.

b. Black-box risks due to a lack of explainability: certain algorithms lead to black-box models where model outputs are very hard to explain. This can lead to issues and mistrust, especially if the model affects human life. Often applies to deep learning.

When the use case does not allow for interpretable models, an unexplainable (black-box) model can be an appropriate choice. Nevertheless, such a lack of explainability can increase risks. Black-box models can amplify other risks, as the lack of explainability can make potential issues harder to be discovered. This includes risks related to concept drift (see 3.4.4 c.), as the impact of concept drift is harder to estimate: it can be unclear which features are most relevant to the system and how changes in these features affect the system [Cohen et al., 2021]. The developer does not understand the decision process of the model, and risks of adversarial attacks increase with black-box models [Cohen et al., 2021]. However, novel techniques like explainable AI (XAI) allow insights into black-box models with the potential to reduce their risks [Arrieta et al., 2020].

c. Unethical or inappropriate feature choices: unethical model features like gender or ethnicity should be explicitly excluded from model learning to reduce bias risks.

Risks from unethical or inappropriate feature choices are directly related to a lack of the best practice to "prevent discriminatory data attributes used as model features" [Serban et al., 2021]. Therefore, the adherence to the practice is already quantified in the benchmarking part. However, we want to give the user the opportunity to elaborate on the concrete issue and implications by including this risk in the structured risk assessment.

d. Other model-related risks: no description provided. Within the framework this section is meant to stimulate the user to think about other risks related to the model.

3.4.6 Area 6. Process and validation-related risks

Similar to traditional software development, risks for AI development can be reduced by appropriate and standardised processes. Machine learning requires different quality assurance than traditional software development due to the data-driven nature [Batarseh et al., 2021].

a. Lack of process automation and formalization: the more manual effort the model development requires, the more resources are necessary. Furthermore, the development process should be formalized to reduce variability and human error.

This risk relates to several automation-related best practices [Serban et al., 2020] and generalizes them to automation and formalization of processes in general.

b. Inappropriate evaluation metrics: training and production performance should be represented by an appropriate metric depending on the use case.

Tao et al. list several examples of quality parameters in AI applications, like accuracy, system stability, and timeliness. Furthermore, the authors imply that the relevant quality attributes depend on the use case (e.g. image recognition) [Tao et al., 2019]. This includes a consideration of the severeness of false positives compared to false negatives. The risk from inappropriate evaluation metrics can also be seen as a generalization of the best practice to "capture the training objective in a metric that is easy to measure and understand" [Serban et al., 2020].

c. Lack of monitoring in production: production performance should be continuously monitored to intervene if performance degrades.

This risk originates from a lack of the best practice to "continuously monitor the behaviour of deployed models" [Serban et al., 2020]. By including the risk in the risk assessment part of the framework, the user is encouraged to consider the possible consequences of such a lack of monitoring.

d. No fallback in case of faulty behavior: depending on the importance of reliability requirements, fallbacks should be in place if model outputs are inconclusive or not available.

In critical systems that could heavily affect human lives if the system fails or malfunctions, missing fallback processes can result in irreparable damage. The European Union (EU) recommends having fallback procedures for such errors in place. Examples given by the EU are simpler rule-based fallback systems or human-made decisions [European Commission, 2019a].

e. Other process-related risks: within the framework, this section is meant to stimulate the user to think about other process-related risks.

3.4.7 Area 7. Risks from unknown or insufficient generalization

Generalization describes how well a model works for different data than the training data. Poor generalization limits the environments in which a model can be used. Furthermore, a model might show unexpected behaviour in new environments and leads to faulty decisions.

a. Intended use case for model or data unclear: it should be documented for which use case the data and model were originally intended for and for which use cases they are tested and expected to work.

Ideally, a model card and data sheet, as proposed by Raji et al. were available

[Raji et al., 2020]. However, Due Diligence cannot rely on the existence of such documents. If the intended use and limitations of the data set and model are unknown, the risk of malfunctioning increases. This is especially true if the investment thesis intends a new use case for the data or model (see next risk).

b. System generalizes poorly for future use cases: if a model only works in a very specific environment, different future use cases may be difficult to implement. This could reduce the value of the AI component.

The potential impact of this risk depends heavily on the investment thesis. If new use cases are likely, then the user should evaluate how well the model can generalize or how much work could be required to adopt the model (e.g. transfer learning).

c. Other risks related to generalization: within the framework, this section is meant to stimulate the user to think about other risks related to generalization.

3.4.8 Area 8. Risks from insufficient skills or knowledge

If the development team lacks fundamental skills or knowledge, issues might be undetected, and future maintenance and development might be hindered.

For this area, no sub-risks were defined. The risk area was inspired by semi-structured interviews and a proprietary scoring model for data analytics at EY.

3.4.9 Area 9. Regulatory and reputational risks

AI applications are expected to increasingly be regulated to protect citizens from harm [European Commission, 2019b]. Compliance with current and future regulations is vital for the use of AI. Furthermore, unethical practices or a general mistrust of AI could lead to reputational damage.

a. Compliance with relevant privacy regulations: machine learning applications should respect privacy regulations during the use of the software and during the training, especially if sensitive information Is used during the training. The model must not reveal sensitive information to unauthorized parties.

For companies operating within the European Union the General Data Protection Regulation (GDPR) is highly relevant, and AI systems should comply to the regulation. Even though the GDPR does not only apply to AI systems, it is especially relevant in machine learning due to its data-driven nature.

b. Compliance with future regulations: AI development should be informed about possible future regulations (e.g. EU AI Act) to ensure future compliance. Otherwise, unexpected changes might be necessary.

The EU AI Act is a proposal for “harmonised rules on artificial intelligence” within the European Union [European Commission, 2021]. Current AI-based systems might require changes to adhere to the EU AI Act, which could lead to unexpected costs after a merger or acquisition.

c. Mistrust or reputational damage: Increasing advances in the use of AI are closely followed by the public. General developments and company-specific actions could cause mistrust and reputational damage depending on public perception.

According to Fast and Horvitz, public perception of AI is rather positive. However, they also found that certain concerns increase, namely a “loss of control of AI, ethical concerns for AI, and the negative impact of AI on work” [Fast and Horvitz, 2017]. A prominent example of reputational damage from improper use of artificial intelligence is the dutch childcare benefits scandal: a system meant to identify fraud around childcare benefits that discriminated citizens with an immigration background, ultimately leading the Dutch government to resign [Holligan, 2021].

d. Other risks related to regulations and reputation: within the framework, this section is meant to stimulate the user to think about other risks related to regulation or public reputation.

3.4.10 Contribution arguments

The structured risk assessment supports the thought process while considering the risks of an AI-based system during Due Diligence. We argue that this structured approach supports the completeness of a risk assessment as the risk areas and their sub-risks can act similar to a checklist. Therefore, it is more likely that the user covers all essential areas during Due Diligence.

4 Evaluation

We evaluated the framework in two steps: first, on a real-world due diligence project and secondly on a fictional due diligence case related to an existing open source software system. The first evaluation step took place after the first design iteration of the framework. Therefore, we only applied the AI criteria and the best practice benchmarking in the real-world project, as the structured risk assessment was only developed after the second design iteration. The second, fictional due diligence case covered only the structural risk assessment due to time constraints. We discuss both evaluation steps (i.e. cases) separately.

4.1 Case study: real-world due diligence case

Due to the research context within a consulting company (see 3.1), we were able to apply the first two parts of this framework on a real-world due diligence project. The target provides educational websites to support students different grades with in studying. The aim of this evaluation is not to judge of the target regarding their machine learning capabilities but rather to obtain insights into the practicality of the AI criteria and best practice benchmarking. Therefore, no conclusions about the target should be made based on the evaluation. The application of interest at the target is a newly developed educational platform that utilizes machine learning to predict the skill level of students on their final examination. The purpose of these prediction is to give teachers and students insights about their progress towards the desired skill level, allowing to increase support if the performance is below the personal goal.

The first two assessment parts were applied on the target application by a consultant under supervision of the author. The framework was primarily applied by the consultant. However, the author was involved in multiple discussions about aspects of the framework and the target for a better understanding of the Due Diligence process and the strength and weaknesses of the framework. The consultant is experienced in Due Diligence, with a focus on technology, commercial, and financial Due Diligence, but rather low experience in machine learning or AI in general. In the following we refer to the consultant as the user for the sake of consistency. We expect that the effectiveness of the framework depends on the experience the consultant has in Due Diligence and with AI.

4.1.1 Evaluation of the AI criteria

The user considered the AI criteria as useful to make a judgement about the extend a system actually exhibits artificial intelligence. However, some criteria were difficult to understand or even wrongly understood by the user. We expect that the criteria are easier to apply the more experience the user has with artificial intelligence. We argue that difficult criteria are less problematic than criteria that are misinterpreted. If the user has difficulties understanding a criteria, he can consult experts or research the related AI field. This still increases the aware of the user on possible implications of the AI system compared to an assessment without the criteria. In other words, if a criteria is difficult to understand because the user lacks knowledge in a specific AI field, than he would not have considered that AI field without the use of such AI criteria. On the other hand, criteria that are misinterpreted are more dangerous: We observed that all indications, including those from misinterpreted criteria, increase the confidence of the user in his understanding of the AI, which could lead to over-confidence regarding their implications.

In consultation with the author, the user identified that the software system includes a machine learning component based on supervised learning.

4.1.2 Evaluation of the best practice benchmark

As the AI criteria correctly indicated a machine learning component, the user could continue with the benchmarking of software engineering best practices for machine learning.

The best practice benchmark required significantly more time due to the large number of best practices compared to the AI criteria. We observed that some of the best practices were very difficult to score due to limited information on the AI component and related processes. Furthermore, some of the practices were difficult to understand for the user or wrongly interpreted. We find that the impact of wrongly interpreted practices is less severe as the benchmarking results are aggregated over multiple practices. One outlier does not necessarily impact the overall picture of the software engineering practices.

The resulting user scores were compared with the benchmark as averages per aggregate (first approach described in 3.3.2). In this particular case, the author had the impression that the user scores were inflated compared to the benchmark scores. It could be possible that engineering teams answers the survey more self-critical than an outsider. However, the consultant find the different dimensions of the scoring by category, effects, and trustworthiness insightful. Even if a user does not report the benchmark results, the insights can support the user in the understanding of strength and weaknesses of the system.

Both, author and user acknowledged the limitation of the best practices to software engineering practices. After applying the two framework parts and a proprietary data analytics assessment, the user and the author derived three relevant risks of the system in a brainstorming session, which led to an extension of the framework with a risk assessment part (see 3.4). The second case study focuses on the evaluation of the structured risk assessment.

4.2 Second case study: Public Eye

We fabricated a fictional due diligence case to evaluate the structured risk assessment. The case is based on an existing open source system in order to increase the validity of the case study: Public Eye is a crowd monitoring system developed by the City of Amsterdam [Gemeende Amsterdam, 2020]. The system monitors public places in Amsterdam via computer vision on camera live streams. The cameras are installed in prominent areas of the city and solely used to predict the crowdedness of a place. The front-end consists of a public and a private part. The public part gives access to a map indicating how busy certain places in Amsterdam are and can be freely accessed [7]. The City of Amsterdam hosts an algorithm register with information of AI applications used by the city, the intention behind the applications, and further technical information [Gemeende Amsterdam, 2020]. We utilized the information about Public Eye in this register for our case study. Furthermore, we provided the participants with screenshots, access to the source code, and selected articles about the system. See appendix A for the document provided to the participants. The participants filled in a technology acceptance questionnaire after completion of the risk assessment (see appendix B).

4.2.1 Background of participants and case study duration

A total of four participants applied the structural risk assessment of this framework on the Public Eye case (see table 3).

	Participant			
	1	2	3	4
Experience in Due Diligence or consulting in years	2.5	3	3	<1
Experience in software development in years	0.5	3	0	5
Time spent applying the framework in minutes	45	40	60	75
Appropriateness of the time spent to examine the system 1: too short, 3: perfectly appropriate, 5: too long	2	3	3	3
Agreement with: ”I am an expert in artificial intelligence” 1: strongly disagree, 5: strongly agree	3	3	3	1

Table 3: background of participants in the Public Eye case study and the time they spent applying the framework.

Two of the participants worked simultaneously on the case with the option to communicate and work together during the risk assessment. For these two participants we also observed their work through a Microsoft Teams meeting. Three of the participants had a work background primarily in Due Diligence or consulting, while one participant had a background primarily in software development. The participants spent between 40 and 75 minutes on the case study. Three participants indicated that the time was appropriate to apply the framework, whilst one participant indicated that the time was too short (see table 3).

[7] Public access to the Public Eye: https://druktebeeld.amsterdam.nl/

4.2.2 Observations

The author observed two participants applying the framework and hold a brief feedback session with the participants after completion of the assessment.

Discussions between the participants seem to improve the outcome of the assessment, as participants can either confirm their viewpoints or improve their understanding in discussions of conflicting opinions. Such teamwork is also more likely in real-world cases as Due Diligence is typically carried out in collaboration by a team [Rietveld, 2022]. Furthermore, we observed that the two participants spend comparably more time on the first risk categories than on the later once due to time pressure. The participants decided to reduce their discussions and focus on populating the notes for relevant risks. However, both participants rated the time as appropriate. After the assessment, the two participants noted that the questions are a good entry-point for discussions. Furthermore, both noted that some questions are difficult to answer, partly due to the nature of the fabricated case: in real-world cases, consultants have the opportunity to request additional information or interviews with the target, which was not possible during this case study.

4.2.3 Technology acceptance

Riemenschneider et al. compared five models to predict the acceptance of new methodologies by software developers [Riemenschneider et al., 2002]. We based our technology acceptance questionnaire on the most important factors identified by Riemenschneider et al. The questions were adopted for the context of a Due Diligence framework. Certain questions were omitted as they were not applicable to the case study (e.g. if the users are required to use the framework in their daily work). In addition, we ask participants to compare the framework to other appropriate frameworks used before, if applicable (two participants worked with an applicable proprietary framework before). Appendix B shows the complete questionnaire.

The survey results show that the participants would use the framework again in future Due Diligence for AI-based applications (see figure 3 on the next page). Overall, the participants agree that the framework increases their productivity (figure 4) and that the framework makes the risk examination easier (figure 5).

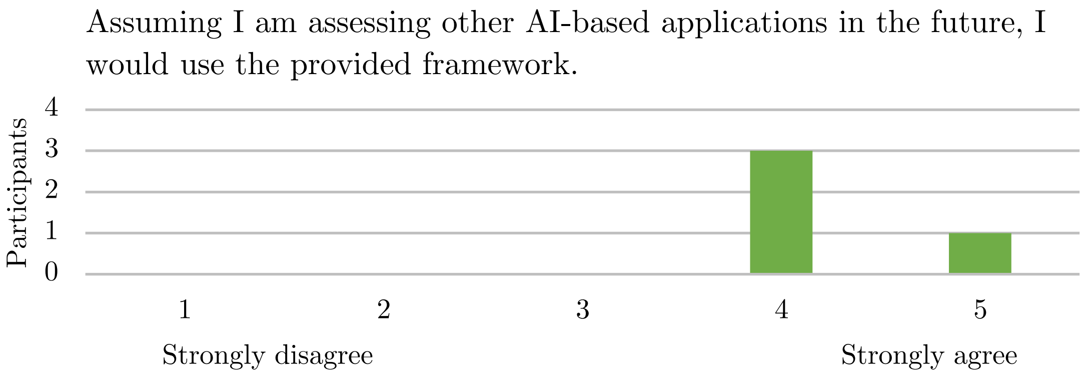


Figure 3: responses of the technology acceptance survey regarding the willingness to apply the framework in the future.

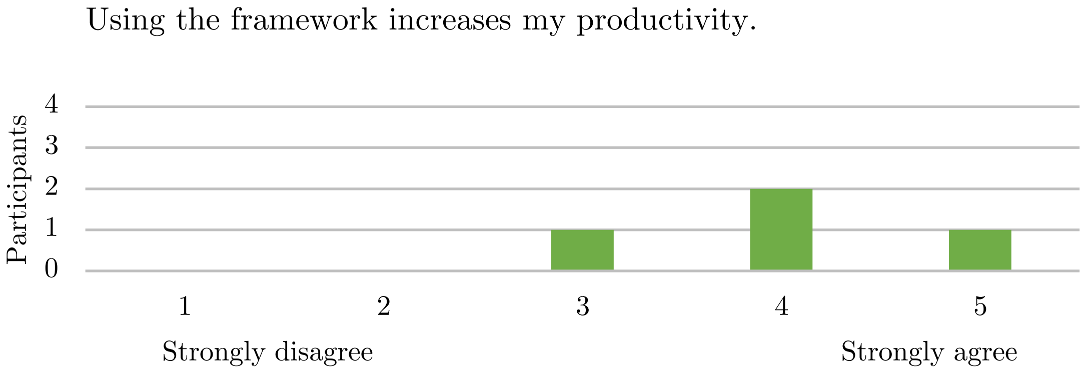


Figure 4: agreement within the technology acceptance survey regarding the productivity increase through the use of the framework

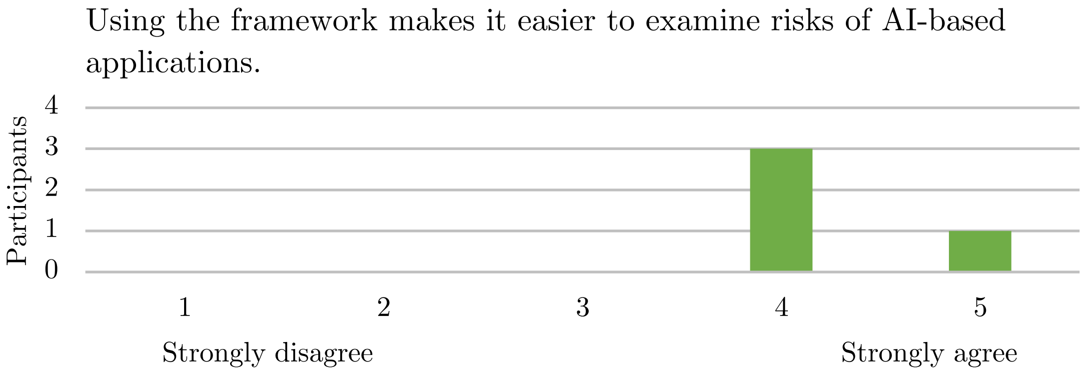


Figure 5: agreement within the technology acceptance survey on weather or not the framework makes the risk assessment easier.

All participants but one indicated that the framework provides sufficient information to apply it 6. Based on oral feedback, it became apparent that the risk categories were rather clear, but the instructions on how to fill in information on the spreadsheet was ambiguous. This concerned a field for notes for each risk, and a drop-down field to indicate weather or not a risk applies. The intention for the "notes" field was to allow the user to elaborate why and how a certain risk applies. This information can later be used to report to the customer. Some participants were unsure if the notes should also be used to give feedback on the framework itself.

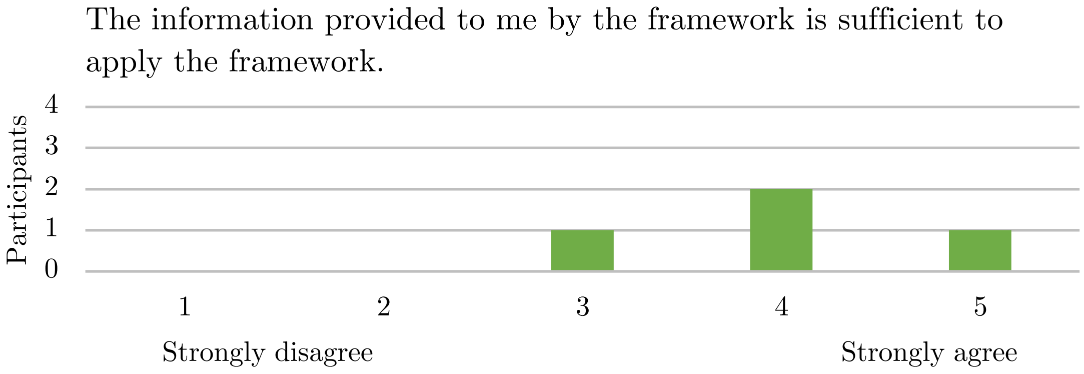


Figure 6: agreement within the technology acceptance survey on weather or not the framework provides sufficient information to apply it.

Due Diligence often follows a semi-standardised process (see section 2.1.2). In addition, we observed that a target often has several applications next to the AI-based systems, which are also assessed. A framework for Due Diligence should be compatible with existing processes, as well as other assessments next to the AI assessment. Therefore, we asked if the framework is compatible with their typical processes. The participants indicated that the framework can be applied within their typical processes (figure 7).

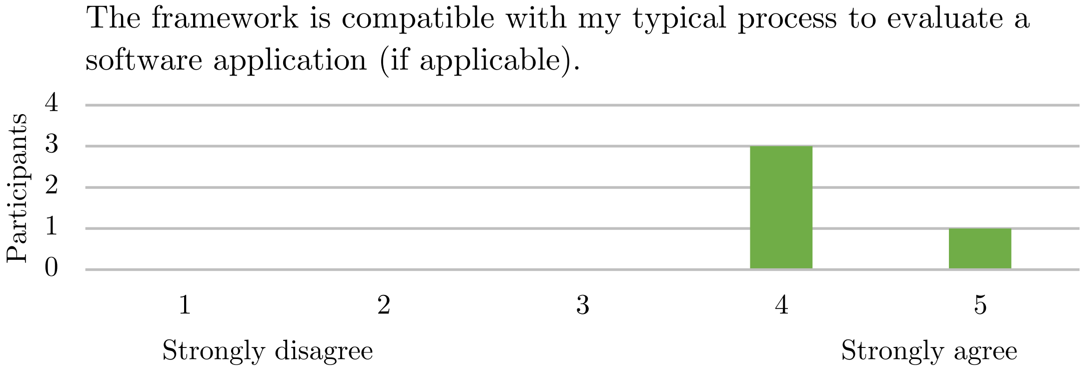


Figure 7: agreement within the technology acceptance survey on weather or not the is compatible with the participant's typical process.

The two participants who used comparable frameworks before agreed that the proposed framework adds value to the existing frameworks. They wrote that "[the framework] complements our model with some new considerations that are not (yet) accounted. As such, combining the two would be sensible". And that the "[risk assessment] generalizes the questions on a higher level, which while providing an overview, creates abstraction".

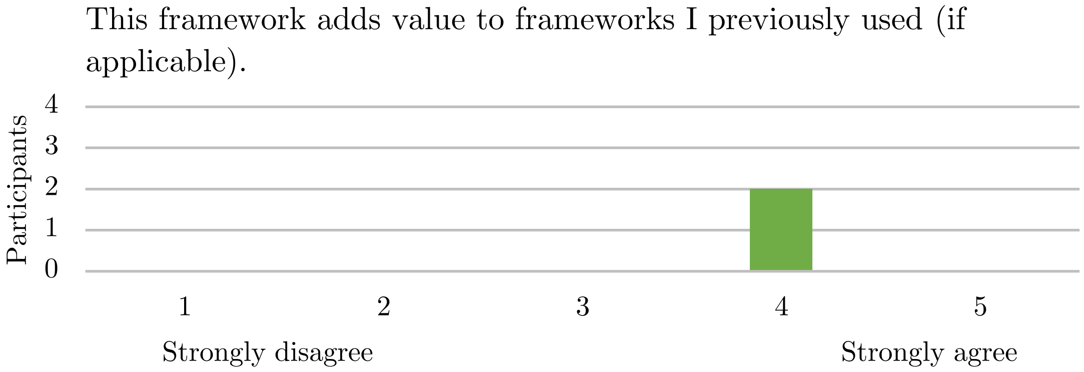


Figure 8: agreement within the technology acceptance survey on weather or not the framework adds value to previously used frameworks.

When asked if the framework can replace previously used frameworks, one participants gave a neutral answer while the other gave a rather positive answer (figure 9). One person noted that the framework could should be used in addition to the their existing framework. Therefore, we conclude that both, the proprietary and the proposed framework, have their own value proposition and do not necessarily compete with each other.

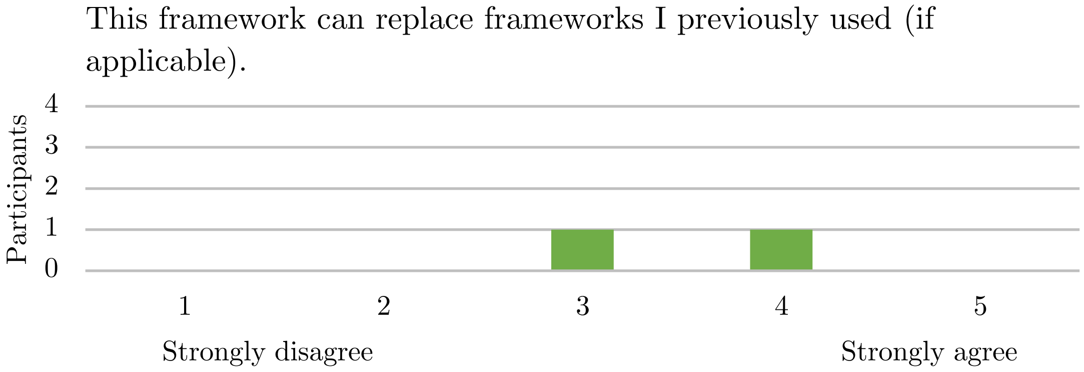


Figure 9: agreement within the technology acceptance survey on weather or not the framework can replace previously used frameworks.

5 Discussion

In this section we discuss the implications of the framework and its evaluation regarding the research questions. Afterwards, we discuss the limitations and threads to validity of our findings.

5.1 RQ 1: How can Due Diligence confirm AI within a software system?

We found that AI is difficult to define due to multiple existing definitions and changing perceptions of AI. Instead of a clear distinction between the existence of AI within a system or the absence of AI, we find that a categorization of a system in different AI techniques is more important for Due Diligence. This enables further research on the techniques used by a system to increase the understanding of their implications for the transaction.

Our proposed AI criteria to identify such techniques were considered useful within the first case study. Nevertheless, in order to minimize errors a certain amount of knowledge of artificial intelligence and machine learning is advised. Our case study indicates that limited knowledge on the described techniques can lead to misinterpretations (see section 4.1.1). Based on our observations, we expect that a cooperative use of the AI criteria with multiple consultants minimizes the risks of misinterpretations. Expert interviews, which are a proven tool in Due Diligence [Rietveld, 2022], are likely to reduce such risks as well.

5.2 RQ 2: What are possible risks of AI-based systems in the context of mergers and acquisitions?

We defined nine risk areas for the Due Diligence of AI-based systems. Eight of which describe possible risks of AI-based systems:

- Mismatch of strategic intent and AI efforts
- Traditional risks for software products
- Data-related risks
- Model-related risks
- Process and validation-related risks
- Risks from unknown or insufficient generalization
- Risks from insufficient skills or knowledge
- Regulatory and ethical risks

Even though we list multiple concrete risks within each area, the list of concrete risks is likely incomplete. AI is a broad field with fast-paced research. Furthermore, most risk areas focus on risks from machine learning, as its data-driven nature differentiates machine learning from traditional software. Nonetheless, we argue that the risks are the most complete set of risks for AI-based systems in mergers and acquisitions, as to our knowledge, no previous research focused on M&As or Due Diligence.

5.3 RQ 3: How can risks of a specific AI-based system be identified and measured during Software Due Diligence?

Software Due Diligence often utilizes supportive tools like checklists, questionnaires, scoring models, static code analysis, and benchmarks [Rietveld, 2022]. We proposed a best practice benchmark and a structured risk assessment to identify which risks apply to a specific AI-based system within a merger or acquisition.

The structured risk assessment supports consultants to identify risks on the eight areas mentioned above. A technology acceptance survey shows positive results, indicating that the assessment is a viable tool to identify risks more easily.

In our experience, risks from AI-based systems cannot be directly measured or quantified. Probabilities for failures and the exact outcome of a risk are unknown. To get closer to a quantified risk we proposed a best practice scoring, allowing to benchmark an application of a target against survey data from machine learning practitioners. From our observation we cannot conclude how accurate the benchmark is. Furthermore, the benchmark is limited to software engineering practices and ignores strategic aspects. Nevertheless, the results of the benchmark can give input to the overall risk assessment, as certain practices are known to result in different effects (e.g. agility and effectiveness). We argue that a lack of effects like agility and effectiveness increases the risks an AI-based system brings to a merger or acquisition.

5.4 Limitations

We acknowledge several limitations of our research. Many aspects of our framework relate to machine learning and are not always directly applicable to other types of artificial intelligence.

Furthermore, the number of case studies is limited to one real-world Due Diligence case and one fabricated case. Due to the diversity of AI-based systems, an evaluation on alternative cases may lead to different results. The author based the second case study on an existing application to improve the validity of the findings.

The observant of the case studies acknowledges that observations during a case study may always be partly subjective. The number of participants was rather small the participants were in a work relationship with the observant. The observant encouraged the participants to provide unbiased and honest feedback.

Although the study was conducted in the Netherlands, we do not expect the findings to vary significantly over different countries. The social context of the study is highly international, the participants of the cases work in an international environment for multinational clients with diverse origins.

Finally, we acknowledge that certain risks may eventually be outdated due to the fast-paced research within the field of AI and especially machine learning. Users of the framework are advised to keep track of recent developments and update the framework if necessary.

6 Conclusion

We researched Due Diligence on AI-based software systems though a review of related research, the development of a Due Diligence framework, two case studies, and observations in the Due Diligence industry. We were particularly interested in how Due Diligence can confirm AI within a software system, what possible risks of AI-based systems are, and how Due Diligence can identify and measure such risks.

We identified nine risk areas within Due Diligence of AI-based systems. Certain risks are valid for Due Diligence in general: many transactions stay behind their expectations, the access to information can be limited and targets may have an inventive to conceal negative aspects of their business. Traditional risks of software products apply to AI-based systems as well, where some aspects like technical debt may be even more common in AI-based systems. Other risks regard the data, model, or processes of the target. Risks for mergers and acquisitions can also stem from unknown or insufficient generalization, insufficient skills or knowledge, and regulatory or ethical aspects.

We developed a three-part framework for Due Diligence on AI-based systems, starting with eight criteria to identify different AI techniques as a starting point for Due Diligence on AI-based systems. A benchmark of machine learning software engineering best practices can quantify some aspects that contribute to the risks of an AI-based system. Furthermore, a structured risk assessment can identify which risks apply to a target in the context of a specific transactions.

We evaluated the proposed framework on two case studies. The results indicate that the framework can support the user in identifying AI techniques of a software system and helps to identify risks of such a system in the context of the transaction. Our observations also indicate, that the best practice benchmarking can support the risk assessment and may be a viable approach to quantify aspects contributing to risks. However, the accuracy of the benchmark is unclear.

For future work we propose an evaluation on more cases. An application of the best practice benchmark through multiple independent teams or people would allow to examine the variability of the benchmark results. Ideally, a longitudinal study would examine weather or not identified risks become apparent after a transaction and which efforts were required to counteract these risks.

7 References

[Angwin, 2001] Angwin, D. (2001). Mergers and acquisitions across European borders: National perspectives on preacquisition due diligence and the use of professional advisers. *Journal of World Business*, 36.

[Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58.

[Batarseh et al., 2021] Batarseh, F. A., Freeman, L., and Huang, C.-H. (2021). A survey on artificial intelligence assurance. *Journal of Big Data*, 8(1):60.

[Bhagwan et al., 2018] Bhagwan, V., Grobbelaar, S. S., and Bam, W. G. (2018). A systematic review of the due diligence stage of mergers and acquisitions: Towards a conceptual framework. *South African Journal of Industrial Engineering*, 29.

[Biggio et al., 2013] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. volume 8190 LNAI.

[Cam et al., 2019] Cam, A., Chui, M., and Hall, B. (2019). Global AI Survey: AI proves its worth, but few scale impact. *McKinsey*.

[Cohen et al., 2021] Cohen, S. N., Snow, D., and Szpruch, L. (2021). Black-box model risk in finance. *SSRN Electronic Journal*.

[European Commission, 2019a] European Commission (2019a). Ethics guidelines for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`.

[European Commission, 2019b] European Commission (2019b). Policy and investment recommendations for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence`.

[European Commission, 2020] European Commission (2020). Sectoral considerations on policy and investment recommendations for trustworthy ai. `https://data.europa.eu/doi/10.2759/733662`.

[European Commission, 2021] European Commission (2021). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[Fast and Horvitz, 2017] Fast, E. and Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

[Felderer and Ramler, 2021] Felderer, M. and Ramler, R. (2021). Quality assurance for ai-based systems: Overview and challenges. volume 404.

[Gemeende Amsterdam, 2020] Gemeende Amsterdam (2020). Public eye - amsterdam algorithm register. https://algoritmeregister.amsterdam.nl/public-eye/.

[Gemeende Amsterdam, 2021] Gemeende Amsterdam (2021). Public eye: an open-source crowd monitoring solution. https://www.amsterdam.nl/innovatie/mobiliteit/public-eye/.

[Gemeende Amsterdam, 2022] Gemeende Amsterdam (2022). Druktebeeld - een beeld van de drukte in regio amsterdam. https://druktebeeld.amsterdam.nl/.

[Ghahramani, 2004] Ghahramani, Z. (2004). *Unsupervised Learning*, pages 72–112. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Gomes et al., 2013] Gomes, E., Angwin, D. N., Weber, Y., and Tarba, S. Y. (2013). Critical Success Factors through the Mergers and Acquisitions Process: Revealing Pre- and Post-M&A Connections for Improved Performance. *Thunderbird International Business Review*, 55.

[Groot et al., 2012] Groot, J. D., Nugroho, A., Bäck, T., and Visser, J. (2012). What is the value of your software? *2012 3rd International Workshop on Managing Technical Debt, MTD 2012 - Proceedings.*

[Haller, 2021] Haller, A. (2021). Global M&A Report 2021 (Bain & Company): The Unsurprising Boom in Technology M&A.

[Happy and Routray, 2015] Happy, S. L. and Routray, A. (2015). Robust facial expression classification using shape and appearance features. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–5.

[Hayes-Roth, 1985] Hayes-Roth, F. (1985). Rule-based systems. *Communications of the ACM*, 28.

[Holligan, 2021] Holligan, A. (2021). Dutch Rutte government resigns over child welfare fraud scandal. `https://www.bbc.com/news/world-europe-55674146`, BBC News.

[IBM, 2020] IBM (2020). Digital acceleration: top technologies driving growth in a time of crisis. `https://www.ibm.com/downloads/cas/MBV83XAY`.

[ISO/IEC 25010, 2011] ISO/IEC 25010 (2011). ISO/IEC 25010:2011, systems and software engineering — systems and software quality requirements and evaluation (square) — system and software quality models.

[Lenarduzzi et al., 2021] Lenarduzzi, V., Lomio, F., Moreschini, S., Taibi, D., and Tamburri, D. A. (2021). Software quality for ai: Where we are now? volume 404.

[McCallum, 2022] McCallum, S. (2022). Clearview AI fined in UK for illegally storing facial images. `https://www.bbc.com/news/technology-61550776`, BBC News.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Muller and Renda, 2020] Muller, C. and Renda, A. (2020). Assessment list for trustworthy artificial intelligence (altai) for self-assessment. `https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment`.

[Pennachin and Goertzel, 2007] Pennachin, C. and Goertzel, B. (2007). *Contemporary Approaches to Artificial General Intelligence*, pages 1–30. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Pfundstein et al., 2021] Pfundstein, M., Sleegers, J., van Arman, T., Groening, D., Groothoff, B., van Woerden, J.-E., de Weerdt, L., and Jongstra, T. (2021). Amsterdam/public-eye. `https://github.com/Amsterdam/public-eye`, GitHub.

[Raji et al., 2020] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing.

[Riemenschneider et al., 2002] Riemenschneider, C. K., Hardgrave, B. C., and Davis, F. D. (2002). Explaining software developer acceptance of methodologies: A comparison of five theoretical models. *IEEE Transactions on Software Engineering*, 28.

[Rietveld, 2022] Rietveld, P. (2022). personal communication.

[Russell and Norvig, 2021] Russell, S. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach, Global Edition 4th*, volume 19.

[Sadiq et al., 2021] Sadiq, R., Safie, N., Hadi, A., and Goudarzi, S. (2021). Artificial intelligence maturity model: a systematic literature review. *PeerJ Computer Science*, 7.

[Schief et al., 2013] Schief, M., Buxmann, P., and Schiereck, D. (2013). Mergers and acquisitions in the software industry: Research results in the area of success determinants. *Business and Information Systems Engineering*, 5.

[Schlimmer and Granger, 1986] Schlimmer, J. C. and Granger, R. (1986). Beyond incremental processing: Tracking concept drift.

[Schulmeyer, 2008] Schulmeyer, G. G. (2008). *Handbook of software quality assurance. 4th edition.* Artech House, Boston.

[Sculley et al., 2015] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., and Dennison, D. (2015). Hidden technical debt in machine learning systems. volume 2015-January.

[Serban et al., 2020] Serban, A., van der Blom, K., Hoos, H., and Visser, J. (2020). Adoption and effects of software engineering best practices in machine learning. In *Proceedings of the 14th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).* ACM.

[Serban et al., 2021] Serban, A., van der Blom, K., Hoos, H., and Visser, J. (2021). Practices for engineering trustworthy machine learning applications.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* The MIT Press, second edition.

[Tao et al., 2019] Tao, C., Gao, J., and Wang, T. (2019). Testing and quality validation for ai software–perspectives, issues, and practices. *IEEE Access*, 7:120164–120175.

[Tom et al., 2013] Tom, E., Aurum, A., and Vidgen, R. (2013). An exploration of technical debt. *Journal of Systems and Software*, 86.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence.

[Turuk and Milovanović, 2020] Turuk, M. and Milovanović, B. M. (2020). Digital due diligence: a complementary perspective to the traditional approach. *International Journal of Contemporary Business and Enterpreneurship*, Vol. I.

[Wang, 2003] Wang, S.-C. (2003). *Artificial Neural Network*, pages 81–100. Springer US, Boston, MA.

[Westenberger et al., 2022] Westenberger, J., Schuler, K., and Schlegel, D. (2022). Failure of ai projects: understanding the critical factors. *Procedia Computer Science*, 196:69–76. International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021.

[Wieringa, 2014] Wieringa, R. J. (2014). *Design science methodology: For information systems and software engineering.*

[Wray, 2021] Wray, S. (2021). Why the city of amsterdam developed its own crowd monitoring technology. `https://cities-today.com/why-the-city-of-amsterdam-developed-its-own-crowd-monitoring-technology/`, CitiesToday.

[Zhang et al., 2016] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. volume 2016-December.

A Public Eye Case Study: Virtual Data Room

“[Public Eye] is a monitoring system with counting cameras and wifi sensors that give insight into numbers and densities of pedestrians. These data are used for strategic, tactical and operational purposes. In the pilot Public Eye Amsterdam, the crowdedness in a number of places in Amsterdam is being mapped with an innovative, open-source part of the data system of [Public Eye]. There are cameras in these locations that are linked to a server of the municipality. An algorithm on the server analyzes how many people are on the images. The information about the number of people detected is forwarded to municipal employees and facility managers, who can use the count to better regulate traffic. The images are not saved nor shown. The algorithm instantly converts the images to anonymous numbers, after which the data are supplied to the [Public Eye]. Residents and visitors of the city can also view the information about the crowdedness on digital information boards at the Marineterrein and online” [Gemeende Amsterdam, 2021].

Product overview

“Amsterdam is a busy city. This can sometimes lead to unsafe situations. By collecting data about the number of pedestrians, it is possible to take measures to control the traffic. This way, the city remains comfortable, accessible and safe for traffic. If a situation becomes unsafe due to excessive congestion, the municipality can intervene. This is done, for example, by placing digital information panels so that people know which routes to take. Or one-way traffic is established” [Gemeende Amsterdam, 2020].

“With the crowd monitoring system 'Public Eye', we map out the crowds in a few places in Amsterdam. At the moment, the system is active on Arena Boulevard, at the Marineterrein and on Dam Square. These locations are equipped with cameras that are linked to a municipality server. On the server, an algorithm analyses how many people appear on the images. The information about the number of people present is sent to municipal employees, who can use the count to better regulate traffic. The images are not shown, only the numbers. Residents and visitors to the city can also view the information on the number of people present, via https://druktebeeld.amsterdam.nl/. At the moment, this is only possible for the Marineterrein location. The ambition is to realise this for all Public Eye locations” [Gemeende Amsterdam, 2020].

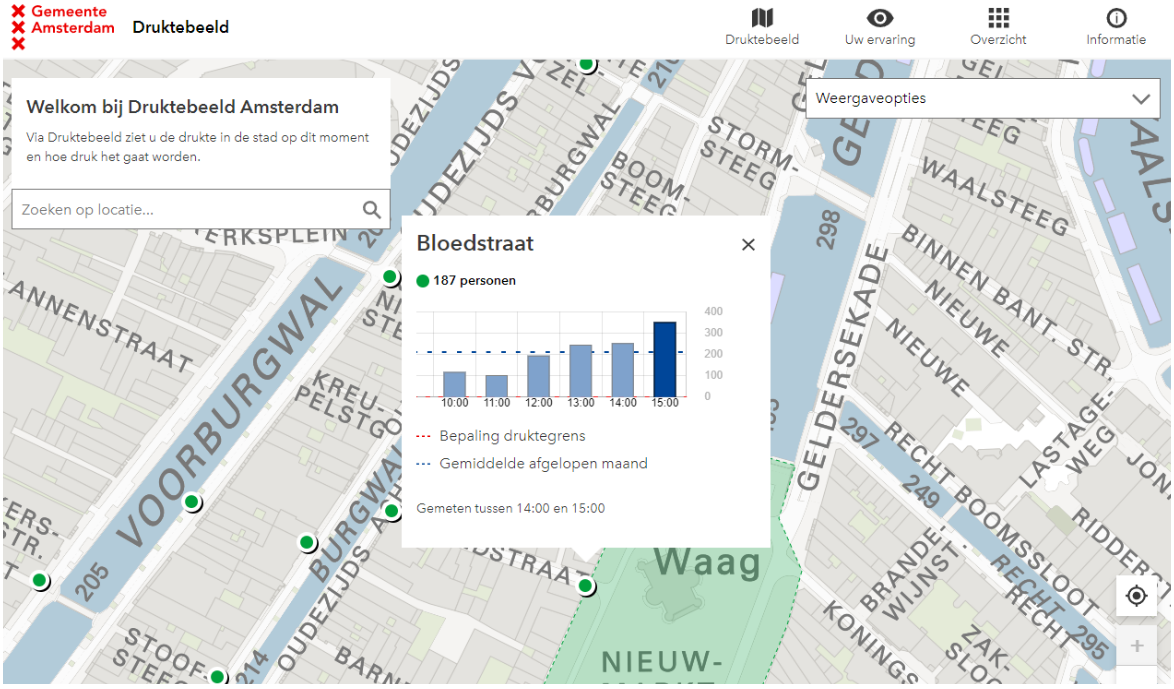
“The video images are immediately deleted as soon as the algorithm has counted the number of people present” [Gemeende Amsterdam, 2020].

“At each new location where Public Eye is placed, a small amount of footage is recorded of which, randomly, about 300 images are annotated for the training of the algorithm. In this way, the crowds at that location can also be properly analysed.

After all, every location is unique and has, for example, a different incidence of light or camera height" [Gemeende Amsterdam, 2020].

Demo

Public access: https://druktebeeld.amsterdam.nl/



Gemeente Amsterdam Druktebeeld

Welkom bij Druktebeeld Amsterdam

Via Druktebeeld ziet u de drukte in de stad op dit moment en hoe druk het gaat worden.

Zoeken op locatie...

Weergaveopties

Bloedstraat

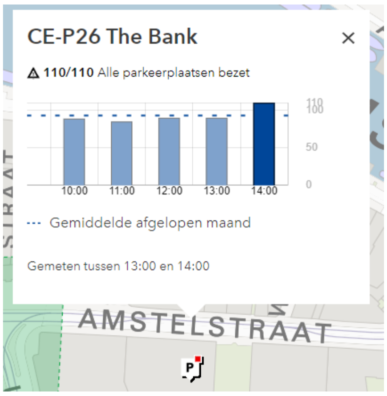
● 187 personen

--- Bepaling druktegrens

--- Gemiddelde afgelopen maand

Gemeten tussen 14:00 en 15:00

Figure II: Screenshot from https://druktebeeld.amsterdam.nl/ [Gemeende Amsterdam, 2022]



CE-P26 The Bank

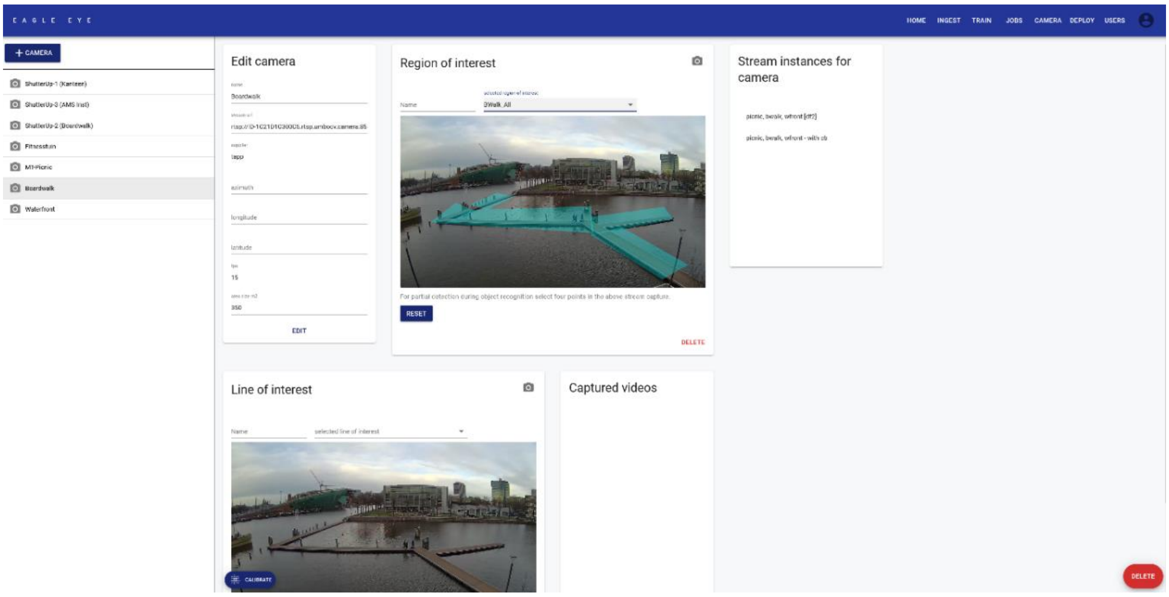
⚠ 110/110 Alle parkeerplaatsen bezet

--- Gemiddelde afgelopen maand

Gemeten tussen 13:00 en 14:00

Figure III: Screenshot from https://druktebeeld.amsterdam.nl/ [Gemeende Amsterdam, 2022]

Backend view to manage camera-related aspects



name: Boardwalk
stream url: rtsp://D-1C21D1C300C5.rtsp.amboov.camera.85
azimuth
longitude
latitude
fps: 15
area size m2: 350

Name — selected region of interest: BWalk_All

For partial detection during object recognition select four points in the above stream capture.

picnic, bwalk, wfront [df2]

picnic, bwalk, wfront - with ob

Name — selected line of interest

Figure IX: "The camera view allows users to manage all aspects related to cameras". [Pfundstein et al., 2021]

Tagging View

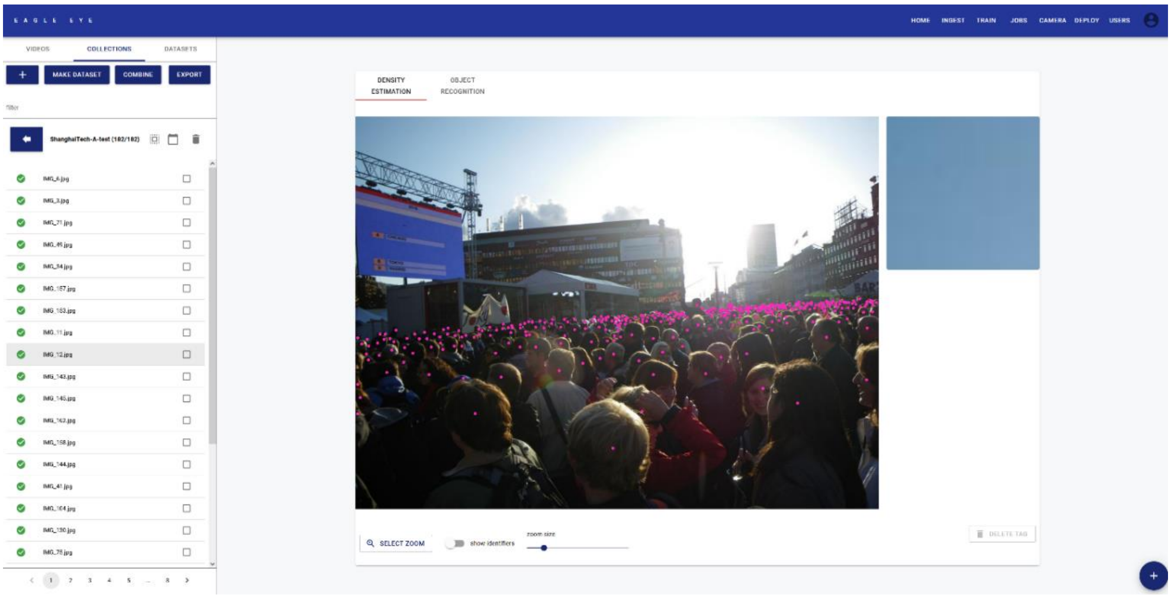


Figure IX: "This view is specialized for dataset creation. Users can tag images that then can be used for training in order to fine-tune the algorithms" [Pfundstein et al., 2021].

Data

Training data: Marineterrein

"With training data, the algorithm 'learns' how many people an image contains. This dataset contains images from four cameras in the Marineterrein area. There are several hundred images per camera. The number of people in the images varies from 0 to approximately 200. The cameras used for these images were at a height of 3 to 15 metres during the data collection at the Marineterrein. In this dataset, we manually annotated where in the image the heads of people are present. These annotations were drawn up in two stages, i.e. each annotation was checked once and adjusted if necessary. In this way, we minimised the chance of errors in annotating. Only a limited number of municipal employees have rights to access this data" [Gemeende Amsterdam, 2020].

Training data: Arena

"This dataset contains images from four cameras in the Arena area. It concerns approximately 300 annotated images per camera. The number of people in the images varies from 0 to 100. The cameras used for these images were suspended around the Amsterdam Arena at a height of 10 to 15 metres during the data collection. In this dataset, we manually annotated where in the image the heads of people are present. These annotations were drawn up in three phases, i.e. each annotation was checked twice and adjusted if necessary. In this way, we minimised the risk of errors during annotation. Only a limited number of municipal employees have rights to access this data" [Gemeende Amsterdam, 2020].

Training data: Dam dataset

"This dataset contains about 1000 images of the Dam Square in Amsterdam. All these images were taken from the same location at the same angle. They are "stitched" images: the images of four different cameras have been merged into a single image. These images show between 0 and 200 people, and the circumstances are always very different. Think of: weather conditions, light, time of day, reflections in the lens due to sunlight, et cetera. In these images, we have manually indicated (or "annotated") where in the image the heads of people are present. These annotations were prepared in three stages, i.e. each annotation was checked twice and adjusted if necessary. In this way, we minimised the chance of errors in annotating. To measure how busy it is, it is of course not necessary to know who is on the images, it is sufficient to know how many people are on the images. Only a limited number of municipal employees have rights to access this data" [Gemeende Amsterdam, 2020].

Training data: Shanghaitech Crowd Counting

Dataset: https://www.kaggle.com/datasets/tthien/shanghaitech [Zhang et al., 2016]

"Shanghaitech Part A: This dataset contains 482 images of large groups of people (average 501.4 per image). These images were collected randomly from the internet.

This set contains annotations indicating the locations of the heads of the people in the image" [Gemeende Amsterdam, 2020].

"Shanghaitech Part B: This dataset contains 716 images of groups of people (average 123.6 per image) captured by various cameras in the city of Shanghai, with different angles of view. This set contains annotations that show the locations of the heads of the people in the image" [Gemeende Amsterdam, 2020].

"The Municipality of Amsterdam did not collect the images of the Shanghaitech Crowd Counting dataset itself. They are freely available via the internet. This dataset is only used for training purposes" [Gemeende Amsterdam, 2020].

Other data involved and the process

"At the moment, the system is active on Arena Boulevard, on the Marineterrein and on Dam Square. These places have cameras that are linked to a server of the municipality. On the server, an algorithm analyses how many people appear on the images. The video images are erased immediately after analysis" [Gemeende Amsterdam, 2020].

Data processing

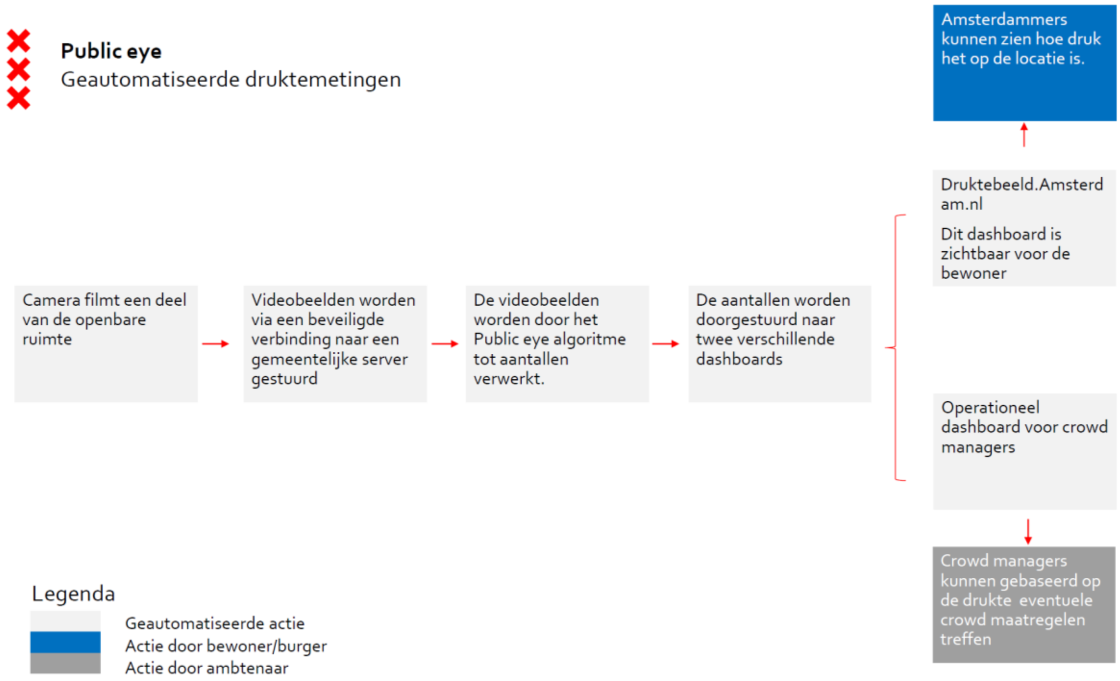

Public eye
Geautomatiseerde druktemetingen
Camera filmt een deel van de openbare ruimte
Videobeelden worden via een beveiligde verbinding naar een gemeentelijke server gestuurd
De videobeelden worden door het Public eye algoritme tot aantallen verwerkt.
De aantallen worden doorgestuurd naar twee verschillende dashboards
Druktebeeld.Amsterdam.nl
Dit dashboard is zichtbaar voor de bewoner
Amsterdammers kunnen zien hoe druk het op de locatie is.
Operationeel dashboard voor crowd managers
Crowd managers kunnen gebaseerd op de drukte eventuele crowd maatregelen treffen
Legenda
Geautomatiseerde actie
Actie door bewoner/burger
Actie door ambtenaar

Figure X: system architecture of Public Eye [Gemeende Amsterdam, 2020].

Model architecture

"A camera takes video images of a certain area. The video images are sent - secured by end-to-end encryption - to a local server. The algorithm analyses how many people are in the images. That number is sent to an overview page (dashboard) for the operational staff of the municipality, so that they have an accurate picture of how busy they are at that moment. In addition, the number is shown on https://druktebeeld.amsterdam.nl/. At the moment, only the busyness at the Marineterrein location is shown, in the future also for the other Public Eye locations. The video images do not leave the server and are not stored. A very limited number of images are only stored for training purposes, these are encrypted (see 'Data')" [Gemeende Amsterdam, 2020].

"Personal data is processed in accordance with the applicable laws and regulations (AVG) and the Transparency Guideline (TADA). A specific privacy statement accompanies this project. The locations and functions of the cameras are included in the camera register of the municipality of Amsterdam" [Gemeende Amsterdam, 2020].

Model performance

"The algorithm must have an accuracy of about 70 percent in order to obtain relevant insights for traffic control. In practice, the algorithm achieves an accuracy of approximately 90%. We derive this from the training images" [Gemeende Amsterdam, 2020].

"In addition to the operational operation, there is constant innovation in this project. We are constantly looking for new functionalities that can improve the system:" [Gemeende Amsterdam, 2020]

"One ambition is to make the system even more privacy-friendly, and we are doing this by adding a model, which allows the algorithm to train with fewer images per camera. This is called the ViCCT model" [Gemeende Amsterdam, 2020].

"We want to perform the analysis on the sensor (the "on edge" technique / edge computing)" [Gemeende Amsterdam, 2020].

Privacy

Privacy concerns were also an important factor in developing Public Eye. "The system decreases the need to monitor camera feeds featuring individuals by turning them into numbers and heatmaps. The camera footage is not saved or recorded, apart from a small number of images for algorithm training, and is processed on a city-owned encrypted network" [Wray, 2021].

Anti-discrimination efforts

"The algorithm is trained with all kinds of images without taking into account the appearance of the people in those images. The algorithm only counts the number of heads" [Gemeende Amsterdam, 2020].

Supervision

“Based on the training data, the quality and accuracy of the algorithm is periodically evaluated by a small number of municipal employees who have permission to view the images. They check whether the algorithm correctly recognizes people as human beings” [Gemeende Amsterdam, 2020].

Risks and prevention

“The video images used by Public Eye are deleted once the algorithm has counted the number of people present. Only for training the model a small number of video images are kept (roughly 300 images p. location)” [Gemeende Amsterdam, 2020].

“The images are on the municipal infrastructure that complies with the Baseline Information Security Government [...]. If the un-anonymised images do end up in the wrong hands, the risk of a breach of privacy is relatively low: the camera is at such a high altitude that it is difficult to recognise people on the images. In addition, data minimisation is applied: the cameras in the ArenA area are only switched on from two hours before an event until the event is finished. At other times, the cameras of Public Eye in the ArenA area are switched off. Work is underway to ensure that the Public Eye cameras at the other locations are also switched off at times when they are not necessary, such as at night” [Gemeende Amsterdam, 2020].

“In order to inform Amsterdammers as accurately as possible, each camera has a sticker with a unique ID code, so that you can use maps.amsterdam.nl/privacy to find out what the camera is for. In this case, it is only telephoto cameras. The privacy policy of the municipality of Amsterdam can also be found on this website: https://www.amsterdam.nl/privacy/” [Gemeende Amsterdam, 2020].

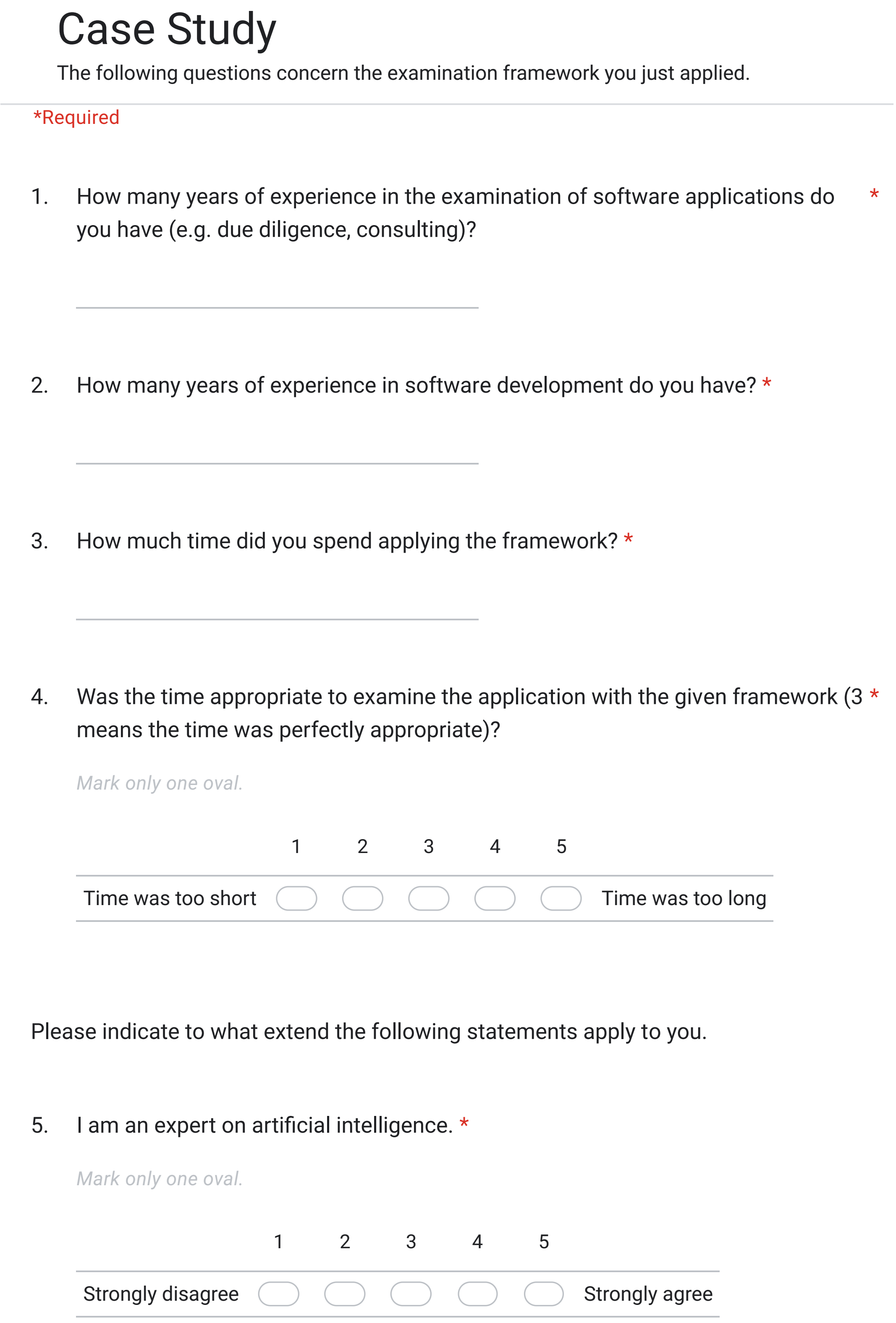
Project planning

“The CMSA was used for the first time during SAIL 2015 and is now permanently posted in de Wallen area, at the Central Station bus platform and near the ferries. At this moment, the Public Eye system is being used and tested at the Johan Cruijff Boulevard, Marineterrein and Plein '40 - '45. Municipality and partners cooperate to continuously improve and further develop the system. For instance, Public Eye is working on the further development of instant analysis of images within the camera, so the data can stay inside the camera. Also, an events calendar is being integrated in the system, so the cameras on the Johan Cruijff Boulevard will only be active during events. Furthermore, in addition to the digital information boards and the crowdedness information webpage, new channels are explored to better inform residents and visitors about the expected crowdedness in the city. This way they will be able to make a well-informed decision about visiting hotspots in the city” [Gemeende Amsterdam, 2020].

Technology

Python for machine learning. Typescript & HTML for frontend and REST backend. “State of the Art performance that outperforms black box software of commercial vendors. Fully functional Web-UI that simplifies all the tedious tasks. Integrates with every RTSP camera in a matter of minutes. Pre-trained ViCCT Transformer for out of the box deployment. Pre-trained Yolov5 based Social Distancing feature. Pre-trained LoID algorithm for Counting Line Crossings and Directions. Multiple user-defined Regions-of-Interests per camera. HTTP(s) Webhooks to integrate with your ESB or Data brokers. Integrated manual with explanations for each section of the UI. User control management system with fine-grained permission system. Data management system for video and image data. Intuitive tagging system for rapid dataset creation. Experiment framework for keeping tracks of training sessions and outcomes. Model database” [Pfundstein et al., 2021].

B Technology acceptance questionnaire



Case Study

The following questions concern the examination framework you just applied.

*Required

1. How many years of experience in the examination of software applications do you have (e.g. due diligence, consulting)? *

2. How many years of experience in software development do you have? *

3. How much time did you spend applying the framework? *

4. Was the time appropriate to examine the application with the given framework (3 means the time was perfectly appropriate)? *

Mark only one oval.

	1	2	3	4	5	
Time was too short	○	○	○	○	○	Time was too long

Please indicate to what extend the following statements apply to you.

5. I am an expert on artificial intelligence. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

6. Assuming I am assessing other AI-based applications in the future, I would use the provided framework.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. Using the framework increases my productivity. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. Using the framework makes it easier to examine risks of AI-based applications. *

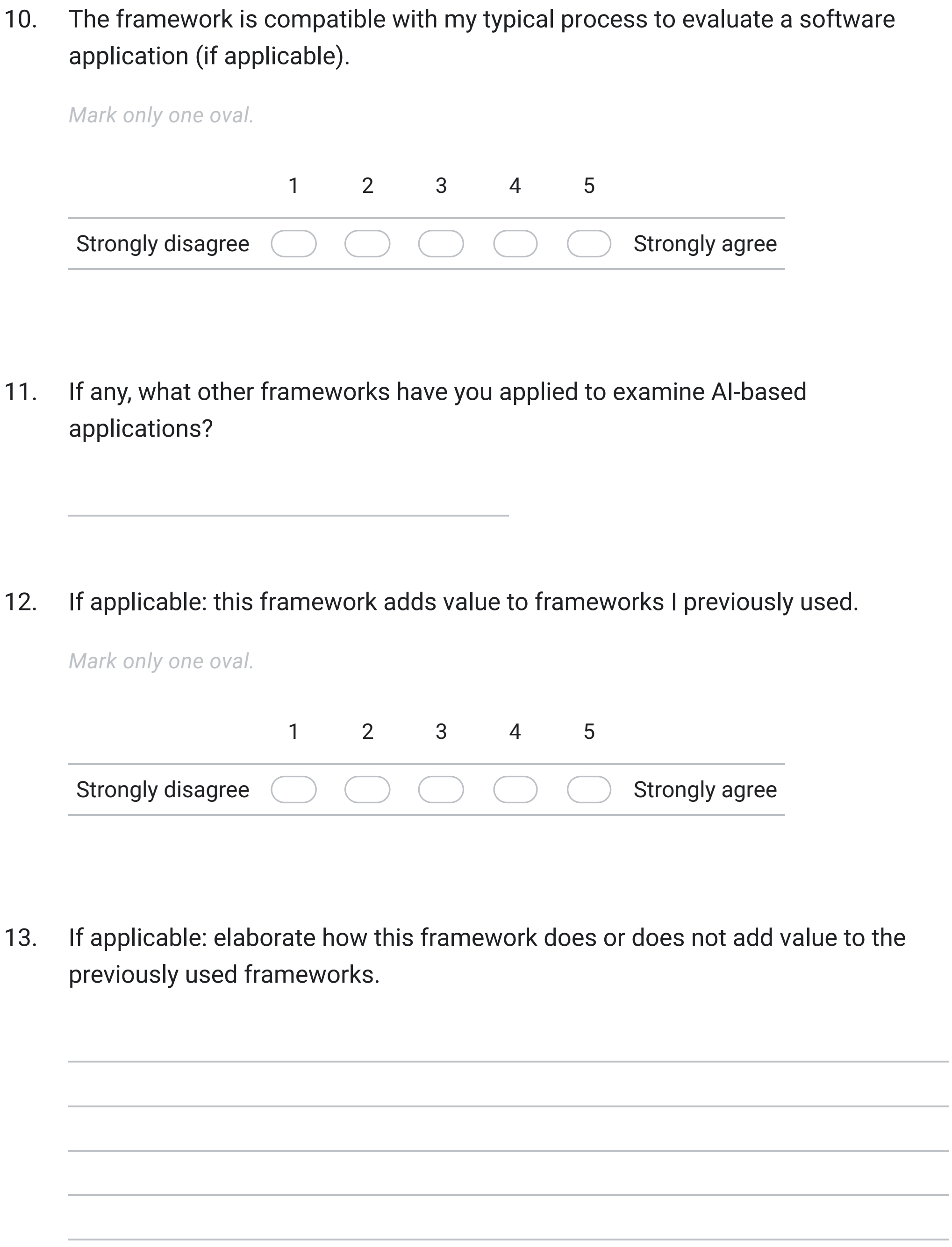
Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. The information provided to me by the framework is sufficient to apply the framework. *

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree



10. The framework is compatible with my typical process to evaluate a software application (if applicable).

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

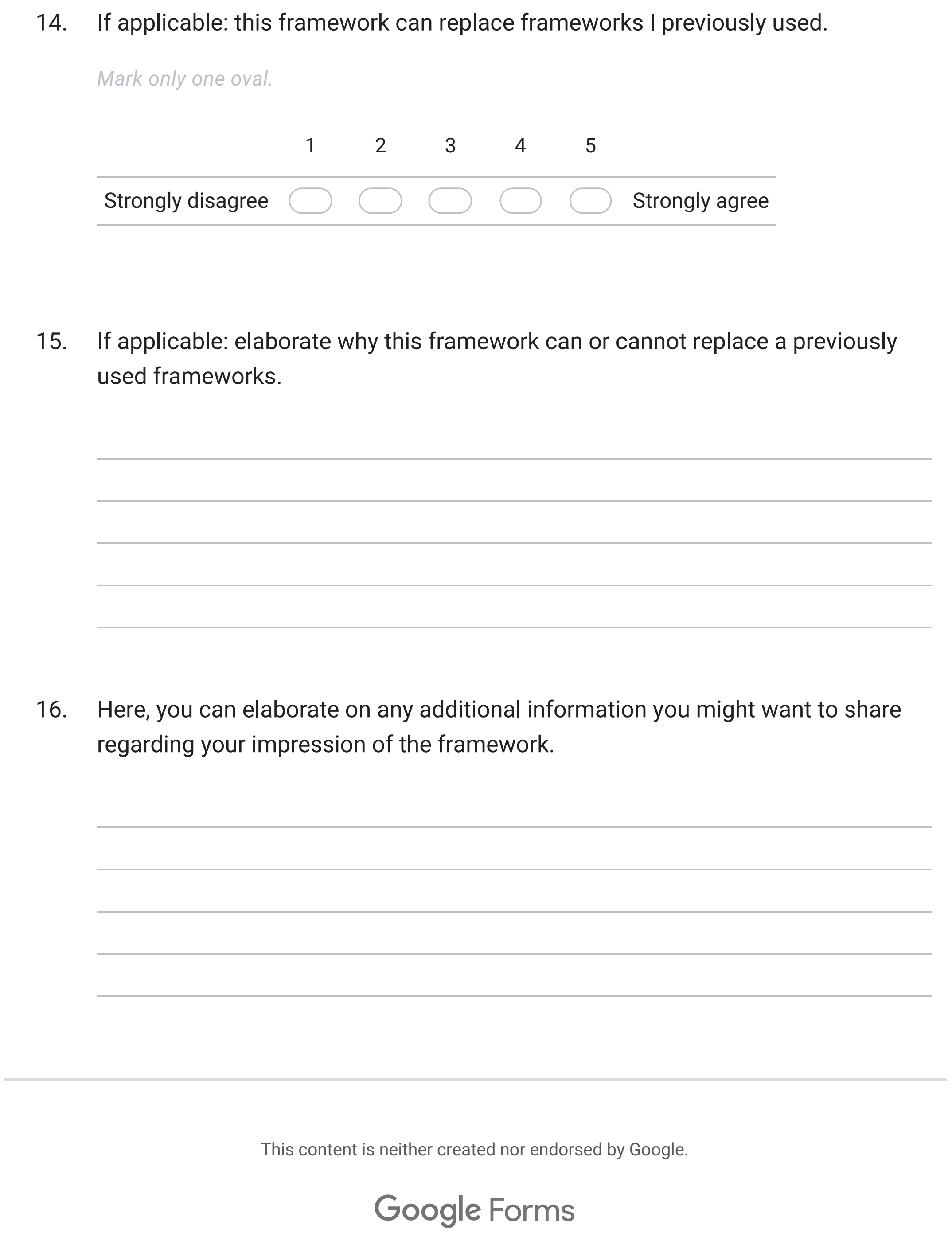
11. If any, what other frameworks have you applied to examine AI-based applications?

12. If applicable: this framework adds value to frameworks I previously used.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

13. If applicable: elaborate how this framework does or does not add value to the previously used frameworks.



14. If applicable: this framework can replace frameworks I previously used.

Mark only one oval.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

15. If applicable: elaborate why this framework can or cannot replace a previously used frameworks.

16. Here, you can elaborate on any additional information you might want to share regarding your impression of the framework.

This content is neither created nor endorsed by Google.

Google Forms