Tuomo Sipola
Janne Alatalo
Monika Wolfmayr
Tero Kokkonen *Editors*

Artificial Intelligence for Security

Enhancing Protection in a Changing World

Springer

Artificial Intelligence for Security

Tuomo Sipola • Janne Alatalo •
Monika Wolfmayr • Tero Kokkonen
Editors

Artificial Intelligence for Security

Enhancing Protection in a Changing World

Springer

Editors
Tuomo Sipola
Jamk University of Applied Sciences
Jyväskylä, Finland

Janne Alatalo
Jamk University of Applied Sciences
Jyväskylä, Finland

Monika Wolfmayr
Jamk University of Applied Sciences
Jyväskylä, Finland

Tero Kokkonen
Jamk University of Applied Sciences
Jyväskylä, Finland

ISBN 978-3-031-57451-1 ISBN 978-3-031-57452-8 (eBook)
https://doi.org/10.1007/978-3-031-57452-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial intelligence (AI) has seen unprecedented revival in the public eye during the last years. Recent advances, such as AI-based image generation and large language models (e.g., ChatGPT), have demonstrated the potential of this technology. The use of AI technologies to protect the world we live in is topical as new threats emerge. After our previous book, *Artificial Intelligence and Cybersecurity: Theory and Applications* (Springer 2023), we were left with the feeling that not everything had been said about the topic. Consequently, we gathered a group of international experts who could write about the role of AI in the changing world.

This book is divided into three parts, concerning methodological fundamentals, critical infrastructure protection, and anomaly detection. This division emerged from the contents of the chapters that we received. It was quite natural to have chapters about protection of critical infrastructure and about anomaly detection methods for digitalized solutions.

The first part is about methodological fundamentals of artificial intelligence, especially within the scope of security. Adrowitzer et al. provide a blueprint towards a safe world with AI and its development and application. Holmström et al. discuss the use of AI from the point of view of organizational and managerial cybersecurity while evaluating its silver bullet status in the hype discourse. After these general introductions to the topic, we turn to deeper investigations about AI methodology. Data are the raw material of most AI work, and protecting the privacy of the individuals whose data are used is an important concern. Kilpala and Kärkkäinen present a review of ways to evaluate differential privacy models. Furthermore, Van Gerwen et al. discuss the challenges of explainable AI in the context of threat intelligence. Jansevskis and Osis, on the other hand, explain knowledge discovery frameworks and how to include security considerations into them. This relates back to the importance of data and knowledge extraction. This part is concluded by the chapter written by Glazunov and Zarras, which considers the robustness of deep learning. This chapter focuses on technical details, presenting multiple attacks and their significance.

The second part is about the use of artificial intelligence for critical infrastructure protection. As an introduction to the topic, Nweke and Yayilgan explore the use

of AI for the protection of cyber-physical systems. After this, we continue with domain-specific studies. As a first example, Rasmus studies the use of AI tools for small enterprises in the context of security. Another domain is covered by Kiviharju in the chapter about cybersecurity for logistics. Energy, communications, and healthcare are perhaps the most well-known examples of critical infrastructure. Consequently, Martinelli et al. continue with a study about protecting smart grids. Zolotukhin et al. discuss the protection of mobile networks against adversarial examples. Finally, Jonske et al. introduce the reader to the healthcare domain in their chapter about teaching machine learning with medical data.

The last part of the book includes three chapters about artificial intelligence for anomaly detection in various scenarios. Falzone et al. emphasize the importance of automated monitoring of log data and demonstrate the use of AI to detect anomalies in real time. Shahrivar and Millar detect attacks in large-scale event data using machine learning. The book is concluded by Alqarni and Azim who use deep learning to detect anomalies in Internet of Things (IoT) networks.

This book is useful to professionals who are interested in using artificial intelligence for security purposes. It will also be helpful to those who have concerns about its use in the various industry domains. Understanding latest advancements in this field should be useful to those who want to understand modern cybersecurity in detail, and especially to experts in the field, who want to follow research and the latest trends.

Two conflicts of interest should be disclosed. First, Kai Rasmus is supervised in his PhD studies by one of the editors, Tero Kokkonen. Second, Mansour Alqarni's place of affiliation, Fanshawe College, has commercial co-operation related to the cyber range at the editors' institution. These chapters have undergone the same editorial process as all the other chapters in this book.

We would like to thank the authors for sacrificing their time to provide our book with interesting and topical chapters. Without their willingness and valuable contributions, this book would not have become a reality. We wish to extend our acknowledgments to the reviewers who have ensured the relevance and quality of the chapters. The review committee is listed in the front matter, except for the reviewers who wished to remain anonymous.

New, rapidly developing technologies present us with new challenges. Artificial intelligence does not differ in this regard, and the security aspects of the changes it brings should be noticed, so that we can build more secure systems. We hope this book provides the reader unique perspectives to enhancing protection in these circumstances.

Jyväskylä, Finland
November, 2023

Tuomo Sipola
Janne Alatalo
Monika Wolfmayr
Tero Kokkonen

Contents

Contributors

Alexander Adrowitzer Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria

Daniel Innala Ahlmark Luleå University of Technology, Luleå, Sweden

Julian Alff Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Mansour Alqarni Fanshawe College, London, ON, Canada
Ontario Tech University, Oshawa, ON, Canada

Simon Andersson Luleå University of Technology, Luleå, Sweden

Akramul Azim Ontario Tech University, Oshawa, ON, Canada

Fin H. Bahnsen Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Oleh Bakumenko Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Marcel Birnbach Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Alexander Buchelt Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria

Jorge Constantino Delft University of Technology, Delft, Netherlands

Amin Dada Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Jan Egger Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Oliver Eigner Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria

Åsa Ericson Luleå University of Technology, Luleå, Sweden

Simone Falzone AEB SE, Stuttgart, Germany

Jana Fragemann Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Sarah van Gerwen Vrije Universiteit, Amsterdam, Netherlands

Misha Glazunov Delft University of Technology, Delft, Netherlands

Gabriele Gühring Esslingen University of Applied Sciences, Esslingen, Germany

Timo Hämäläinen Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

Anton Holmström Luleå University of Technology, Luleå, Sweden

Martins Jansevskis Vidzeme University of Applied Sciences, Valmiera, Latvia

Frederic Jonske Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Benjamin Jung AEB SE, Stuttgart, Germany

Tommi Kärkkäinen Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

Peter Kieseberg Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria

Minna Kilpala Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

Moon-Sung Kim Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Mikko Kiviharju Computer Science Department, Aalto University, Espoo, Finland

Jens Kleesiek Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Bram Klievink Leiden University the Hague, Den Haag, Netherlands

Maxim Kondratenko Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Johannes Kraus Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Johan Lugnet Luleå University of Technology, Luleå, Sweden

Maria Lymbery Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Fabio Martinelli Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy

Fabio Massacci Vrije Universiteit, Amsterdam, Netherlands
University of Trento, Trento, Italy

Francesco Mercaldo Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy
University of Molise, Campobasso, Italy

Stuart Millar Rapid7 LLC, Boston, MA, USA
Institute of AI in Medicine (IKIM), University Medicine Essen, Essen, Germany

Livinus Obiora Nweke Norwegian University of Science and Technology (NTNU), Trondheim, Norway
Noroff Accelerate, Oslo, Norway

Kaspars Osis Vidzeme University of Applied Sciences, Valmiera, Latvia

Kevin Osthues Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Gregor Pavlin Thales Research and Technology, Delft, Netherlands

Kai Rasmus Luode Consulting Oy, Jyväskylä, Finland

Lars Reinike Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Ritten Roothaert Vrije Universiteit, Amsterdam, Netherlands

Antonella Santone University of Molise, Campobasso, Italy

Stefan Schlobach Vrije Universiteit, Amsterdam, Netherlands

Benjamin Schulz Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Pojan Shahrivar Rapid7 LLC, Boston, MA, USA

Fabian Siethoff Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Tobias Simon Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Marlies Temper Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria

Katja Tuma Vrije Universiteit, Amsterdam, Netherlands

Ben Wagner Delft University of Technology, Delft, Netherlands

Joey Wang Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Brecht Weerheijm Leiden University the Hague, Den Haag, Netherlands

Sule Yildirim Yayilgan Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Di Zhang Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

Nils Zhang Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Apostolis Zarras University of Piraeus, Piraeus, Greece

Mikhail Zolotukhin Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

List of Reviewers

António Abreu Instituto Politécnico de Setúbal, Setúbal, Portugal

Tarja Ajo Jamk University of Applied Sciences, Jyväskylä, Finland

Raquel Barreira Instituto Politécnico de Setúbal, Setúbal, Portugal

Svetlana Boudko Norwegian Computing Center, Oslo, Norway

Panagiotis Douris Center for Security Studies (KEMEA), Athens, Greece

Tapio Frantti University of Jyväskylä, Jyväskylä, Finland

David Hästbacka Tampere University, Tampere, Finland

Jari Hautamäki Jamk University of Applied Sciences, Jyväskylä, Finland

Eppu Heilimo Jamk University of Applied Sciences, Jyväskylä, Finland

Pasi Hyytiäinen Jamk University of Applied Sciences, Jyväskylä, Finland

Petri Kannisto Tampere University, Tampere, Finland

Esther Kern Brandenburg Institute for Society and Security, Potsdam, Germany

Lukas Daniel Klausner St. Pölten University of Applied Sciences, St. Pölten, Austria

Gabriela Labres Mallmann University of Jyväskylä, Jyväskylä, Finland

Antti-Jussi Lakanen University of Jyväskylä, Jyväskylä, Finland

Miguel López Instituto Politécnico de Setúbal, Setúbal, Portugal

Antti Mäkelä Jamk University of Applied Sciences, Jyväskylä, Finland

Ilkka Pölönen University of Jyväskylä, Jyväskylä, Finland

Jouni Pöyhönen University of Jyväskylä, Jyväskylä, Finland

Fabi Prezja University of Jyväskylä, Jyväskylä, Finland

Markus Wurzenberger AIT Austrian Institute of Technology, Vienna, Austria

Part I
Methodological Fundamentals of Artificial Intelligence

Safeguarding the Future of Artificial Intelligence: An AI Blueprint

Alexander Adrowitzer, Marlies Temper, Alexander Buchelt, Peter Kieseberg, and Oliver Eigner

1 Introduction

The history of artificial intelligence begins in the 1950s, the first time the term was officially used was in the proposal for the so-called Dartmouth Summer Research Project on Artificial Intelligence in 1956, which was requested by important scientists of the time such as Claude Shannon (the founder of modern information theory), Marvin Minsky, Nathaniel Rochester, and John McCarthy [32]. At that time, the application already contained questions such as "How can a computer be programmed to use a language," something that has now only become possible with language models such as ChatGPT. Even then, people were concerned about the ethical implications of such technology [33, 42, 44, 52], even though it would be decades before the first working applications of artificial intelligence were developed.

Artificial intelligence has found its way into many products of everyday life. Positive examples include vacuum cleaning robots, personal assistants, or assistance systems in cars. Several studies [3, 40] show that such systems bear various concerns, and therefore legal, ethical, and domain-specific considerations have to be made. From this emerges a strong need to safeguard these AI systems from malicious attacks.

When embarking on the development of AI systems, there are established standard processes that can be adopted.

The root for currently used life cycles can be found in data mining processes. A well-known approach is the Knowledge Discovery in Databases (KDD) process [19], which is composed of selecting target data, that is to be analyzed,

A. Adrowitzer (✉) · M. Temper · A. Buchelt · P. Kieseberg · O. Eigner
Department Computer Science and Security, St. Pölten University of Applied Sciences, St. Pölten, Austria
e-mail: alexander.adrowitzer@fhstp.ac.at

T. Sipola et al. (eds.), *Artificial Intelligence for Security*,
https://doi.org/10.1007/978-3-031-57452-8_1

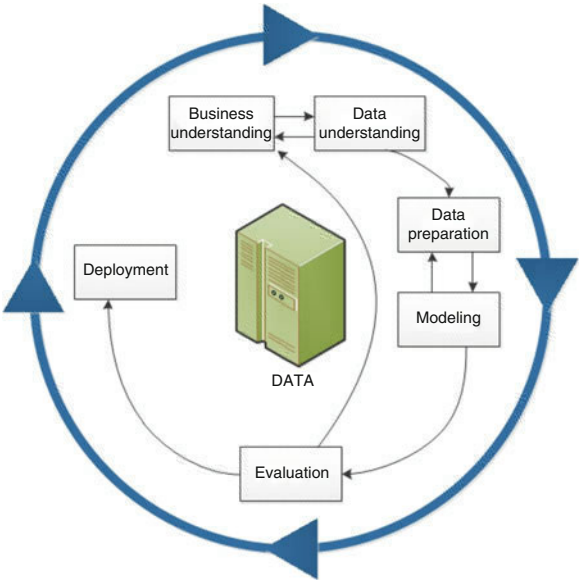

Fig. 1 The six phases of the CRISP-DM lifecycle. Source: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

preprocessing the data and performing necessary cleaning steps, transforming the data, mining existing patterns within the data, and finally interpreting what was found. Patterns like this are important as they provide a standard order of operations for practitioners of the field to adhere to. In using them, certain standards like ethical or moral can be enforced.

The most notable successor to the KDD is the CRoss Industry Standard Process for Data mining (CRISP-DM) [45]. It was developed in 1996 and became a European Union project in 1997 with the leadership of five companies. The methodology was presented in 1999 and published as a data mining guide. In subsequent years, discussions were held for updating the model. CRISP-DM is important due to its widespread adoption and several advantages it offers to the data mining industry. It is the most widely used data mining model, providing a structured and systematic approach to data analysis. It helps address existing challenges in data mining by offering a step-by-step guide for practitioners. Figure 1 shows the original CRISP-DM workflow with its major parts. Its industry, tool, and application neutrality contribute to its success, making it adaptable and applicable across different domains. These features make the model a good basis for developing secure AI applications. We will have a closer look at the different phases of the model now: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

During the business understanding phase, the primary focus is on formulating relevant questions. A skilled data scientist possesses deep knowledge of the application domain they are working in. This domain expertise enables them to identify and articulate questions that can be addressed using analytical methods. While one might assume that finding questions is straightforward, often domain

experts are unaware of existing methods to solve their problems, or they simply go through daily routines without questioning them. When a problem is identified, its initial formulation may prove insufficient and necessitate adjustments.

The data understanding phase involves thoroughly examining the available raw data within an organization. It is crucial to understand both the strengths and limitations of the data in order to effectively work with it. Frequently, data is collected without a specific goal in mind, resulting in potential gaps for addressing certain questions. In such cases, it may be necessary to acquire data from external sources to supplement the existing dataset. If fortunate, the required data might be available as open data; however, there could be associated costs if data needs to be purchased or is not readily accessible. Therefore, the availability of data can directly influence the framing of the question. Data preparation is a time-consuming task for data scientists. Before applying analytical methods to the data, it is essential to ensure that the data is of sufficient quality.

A previous conversion of data into a structured form is often necessary, as this representation of data is particularly suitable for further analyses. Once data is put into the desired form, erroneous, missing, or noisy data must be cleaned using data preprocessing methods. The use of different data mining methods depends on the data category. For example, not every analysis method can work with categorical data. Quantitative data may need to be normalized before modeling methods are applied.

In the modeling phase, different approaches like supervised, unsupervised, or reinforcement learning are applied to the data to find patterns, regularities, or decisions. There are many different algorithms available for this purpose.

The objective of the evaluation phase is to identify the most suitable and valid model. Before applying a model in an organization, it needs to undergo rigorous testing under controlled laboratory conditions. It's important to note that even if a model has demonstrated excellent performance in the laboratory, it does not guarantee similar performance in real-world scenarios, so continuous testing and adapting is necessary.

Deployment refers to the process of implementing the model into regular operation. To effectively utilize models in a production system, they need to be adapted regularly to match the conditions of the operating environment and seamlessly integrate with the existing infrastructure. This may entail significant costs or even the replacement of certain systems. Close collaboration between data scientists and software development teams is crucial in ensuring a successful deployment.

CRISP-DM is a very iterative and promising process, whose phases also alternate with each other. But it lacks of some necessary phases which enables to develop trustworthy AI; for example, it does not inherently take security and risk assessment into account. In 2021, M. Haakman et al. [24] tried to show that AI life cycles need to be revised, because since their inception, the challenges concerning AI have changed and new ones have arisen. After conducting interviews, Haakman et al. propose the addition of a Data collection step, a Documentation step, a Risk Assessment step, and a Monitoring step after deploying a model.

In this chapter, we would like to highlight aspects that would have to be added to the CRISP-DM in order to meet all requirements for the development as well as the deployment of safe AI applications. While safe AI has different definitions, in the context of this chapter, this term is focused on systems that pose minimal risk of failure or discrimination. Therefore, we present a blueprint for AI which, in addition to the phases of the CRSIP-DM, adds missing aspects, regarding explainability and robustness.

The blueprint consists of seven pillars that are necessary for AI systems. These seven pillars must also be taken into account by data scientists, the professionals which are responsible to develop AI systems. A data scientist combines expertise in statistics, mathematics, programming, and domain knowledge to extract valuable insights and knowledge from large and complex datasets. They use various techniques, tools, and algorithms to collect, preprocess, analyze, interpret, and model data in order to solve complex problems and make informed decisions. Additionally, they play a crucial role in developing data-driven strategies and solutions for organizations and are not infrequent responsible for designing, building, and maintaining the infrastructure and systems that enable the storage, processing, and analysis of large volumes of data. The knowledge of necessary steps to provide trustworthy AI is important.

First, in Sect. 2, we consider what influence domain-specific knowledge has on the quality of AI algorithms. In Sect. 3, we will look at the technical aspects; this includes among others the handling of data and the creation of models for machine learning. We will also look at what role structured processes play in this context. Special security aspects are described in Sect. 4, followed by a discussion of ethical aspects Sect. 5. In Sect. 6, we look at the social aspects that are significant in the development of AI. These are mainly the United Nations Sustainable Development Goals (SDG) with a specific focus on environmental aspects. Finally, Sect. 7 concludes this chapter.

2 Domain Aspects

In this section, we give a short discussion in the integration of domain experts into data analysis and discuss some important aspects that are often overlooked when challenging a data project purely based from a data scientists point of view. Since every domain has its own specific merits and peculiarities, introducing domain experts into related data science projects is of the utmost importance, especially when the results ought to be used later on in either commercial products or permanent local installments.

2.1 Integration of Domain Experts

Data scientists may lack expertise in business and domain knowledge. While data science competence can sometimes be a valuable asset in compensating for a lack

of business knowledge, it can also lead to neglecting crucial information that is needed to produce models with relevant business outcomes. Data mining processes should therefore reflect this importance. Currently, the modeling process is abstract and filters out many domain-specific factors that are essential for connecting academic research-based findings with practical, industry-focused problem-solving solutions [6]. Waller et al.[51] describe the importance of domain knowledge in the field of supply chain management. They state that domain knowledge and the analysis of data cannot be separated.

Domain experts, as the name suggests, are well-versed in understanding the data connected to their field of business and can provide valuable insights. Therefore data scientists need to work closely with such domain experts to learn from them. This is the only way to find questions that can be answered with the help of data as well as to navigate around pitfalls.

Having knowledge of innovation methods can also be beneficial for data scientists. Utilizing such techniques, for instance, can aid in identifying questions that have the potential to lead to new data-driven products, business models, or novel fields of application.

We recommend the usage of methods like data canvases [5], persona development, stakeholder interviews, or stakeholder maps [43] as tools that can help data scientists interact with domain experts. In using these tools, data scientists can reach a common ground with domain experts, which breed crucial understanding. First breakout and brainstorming sessions help to identify the required domain experts that need to be integrated, as this is often far from trivial in the setting of a larger company. Thus, the suitability and especially completeness of the people involved must always be challenged critically, especially when departments are suspected to send rather junior, and thus cheap, personal to the respective workshops.

At the same time, data must not be lost sight of. The focus is on acquiring an initial understanding of the available data, its quality, and its suitability for the possible applications at hand. This involves exploring the data, identifying its sources, understanding its structure and format, and assessing any limitations or issues.

This is important for several reasons. Firstly, it helps the data science team gain insights into the characteristics of the data they will be working with. This understanding aids in making informed decisions throughout the project, such as selecting appropriate modeling techniques and determining the feasibility of achieving the desired goals. Secondly, it might give clues for additional features, as well as new strategies for data exploitation in the context of the company. Data scientists must be weary though to not introduce a meaning and usefulness into data sets that quite isn't there, e.g., due to lack of data quality or sample selection bias.

2.2 Providing Trustworthiness and Control

Trustworthiness is considered to be a major important requirement for many data-driven products and services, as it ensures a certain compliance to principles

related to human oversight and security. Nevertheless, trustworthiness in most definitions (see [26] for the definition of the High Level Expert Group HLEG and [49] for the definition by the NIST for two of the most prominent ones) has some pitfalls, as it sometimes incorporates highly nontechnical requirements like ”societal well-being” [26] that might not coincide with functional requirements and typically requires explainability, which is a very strict and demanding feature. Thus, controllable AI [29] could be taken into consideration as an alternative, as it does not impose these requirements onto the AI system but models them rather like we model legal requirements: The operator of the system needs to be able to control that the AI does not diverge too much from the intended mode of operation (in quality of the results, as well as in the way they are achieved) and is able to circumvent the AI in cases this happens. Explainability is not required, i.e., the operator does not need to understand why an AI is or is not working as intended, as long as the overall system is resilient enough to be able to cope with the effects.

2.3 Security and Cyber Resilience

Data understanding also plays a significant role in developing secure data-driven applications. By thoroughly examining the data, potential vulnerabilities and risks can be identified early on. This allows for the implementation of necessary security measures to protect sensitive information and ensure compliance with regulations. A very important topic is cyber resilience. In this concept, related to cyber security, it is assumed that it is impossible to always defend against attacks and that some attackers will get through even the most sophisticated defenses [4]. Thus, a resilient system needs to be able to cope with successful attacks and either recover or change itself in order to be able to continue working as intended. While this is already difficult to achieve in non-AI systems, the missing explainability makes even standard procedures like penetration tests complicated in certain instances of AI, especially when considering self-changing systems like in reinforcement learning. Still, understanding the data might help in uncovering potential biases, errors, or inconsistencies that could impact the reliability and fairness of the resulting models, or, at least, allowing for more informed risk management. Furthermore, sanity checks could be devised that allow for at least some form of detection mechanisms regarding output or model manipulation. The output of such an analysis is often realized as a quality report on the data, describing all data sources and all fields in the data set, as well as information on how control is exerted upon them. Still, securing AI and providing resilient AI systems are a big challenge for many of the very popular AI techniques like deep learning or reinforcement learning and requires much more additional research, which will require knowledge on the domains involved.

2.4 Laws and Regulations

Another vital domain-specific aspect is the adherence to all sorts of norms, standards, laws, and regulations that are in place. While this is more or less standard for common regulations like the GDPR, knowledge on domain-specific regulations needs to be derived from cooperation with the respective experts. This is especially important in case of standards and best practices that are not made into law, as these are very hard to find out about from a pure outside perspective. This especially holds true, when these best practices do not apply for the industry as a whole, but only to a very specific subset of systems. Furthermore, as new systems should be designed to be as future proof as possible, upcoming regulations should be taken into account too. Following, we give a short example on some very important pieces of legislation that will come into effect in the European Union in recent years that will affect the usage of AI. Please keep in mind that this list is not comprehensive.

2.4.1 The AI Act

The most prominent example for novel regulatory development in the area of AI is the AI Act [14]. Its main target lies in providing the rules for a common market for AI-based systems. It not only provides a definition of what is considered to be AI, which had been the topic for a lot of debates, but also classifies the utilization of AI based on risks and application domains involved into four different categories: prohibited, high risk, limited risk, and minimal risk. For each risk class, and especially for high risk AI, guiding principles and rules are defined. At the time of writing this paper, some paragraphs of the AI-Act are still very much under discussion; thus, we omit discussing further details at this point.

2.4.2 The Data Act

The Data Act [16] focuses on manufacturers of smart devices and cloud providers and especially focuses on control over data generated by these devices. The idea behind the Data Act lies in increasing availability and interoperability of nonpersonal information and comes with several requirements for the developers and providers of IoT devices. Firstly, the design of the products must be done in a way to ensure simple real-time access to the collected data generated by the devices, as well as by connected ones. Furthermore, it grants new privileges for the owners of connected products, especially the right to request data holders to share data with a specific third party directly. In addition, it also grants privileges to public bodies, especially the right to request access to the data in cases of emergency.

2.4.3 The Data Governance Act

The main purpose of this act [13] lies in providing a framework for ensuring confident data sharing that can be reused easily from a technical perspective – aiming at providing an increasing amount of data of good quality in order to fuel innovation. This also demands infrastructure setup by the member states in order to facilitate this reuse and defines the duties of so-called *data intermediation services* as well as defines the concept of *data altruism*.

In addition, other software or system-related regulations like the NIS2-directive [15] might have an impact on design and use of a specific AI system, especially in case of critical infrastructures. Furthermore, even company-specific standards and best practices might be in place that require addition attention and need to be taken into consideration too.

2.5 Domain-Specific Peculiarities

Including domain experts is also very important in order to integrate AI-based systems well with existing functionality, especially considering system-specific nonfunctional requirements. As an example, industrial environments typically have long lifespans, which not only make the addition of new hardware and software difficult due to compatibility and performance issues but also can be problematic when introducing standard features for cyber resilience, which in turn makes achieving trustworthiness difficult. As an example, deep package inspection might not be introduced due to the performance overhead of the inspection which would lead to problematic delays in certain arts of an industrial complex. Even more problematic is the standard doctrine of patching, which is hard to do in many industrial environments due to problematic downtimes or the need for recertification [28]. In addition, domain knowledge is extremely important when dealing with real-world data, as this is typically tainted, i.e., there are errors inside the data, either from the data generation/retrieval process or due to problems in the subsequent handling. Thus, in any real-world application, data cleansing [25] is of the utmost importance, which not only requires a lot of domain knowledge but also has the danger of opening a plethora of legal issues [47].

3 Technical Aspects

In this section, we will discuss some selected aspects from the technical perspective, especially focusing on the CRISP-DM model for its structured approach. In addition, we focus on modeling the security and privacy relevant parts and comment on issues that need further reflection apart from the original approach.

3.1 Data Preparation

Data preparation is a vital part in any data-reliant system which is often forgotten or underestimated. This especially holds true for the data cleansing stage, which is not even mentioned in many scientific publications using data, despite the potentially large impact it can have on the actual results. This also holds true for anonymization techniques, which potentially introduce a lot of distortion, e.g., in the case of using generalization-based approaches for reaching k-anonymity [48]. More importantly, it has been shown in [46] that it is not even possible to give good estimations, or even lower/upper boundaries, on the distortion introduced, when executing machine learning algorithms on k-anonymized data sets: First, the actual data quality of the anonymization is largely depending on the actual data precision metrics in use [12], which is currently a largely underdeveloped topic. Second, even when comparing anonymization using the same metrics with different algorithms on the same data set, the distortion differs vastly. Third, cases have been identified where the subsequent machine learning algorithms performed better on anonymized data sets of lesser granularity (i.e., data sets that have been anonymized *stronger*), which is counterintuitive at first glance, but logical, if the anonymization by generalization is seen as a form of pre-clustering. In some singular cases, working on the anonymized sets yielded even better results than working on the original ones, due to outlier removal and pre-clustering effects. Contrary to these results, in many cases, the distortion of strong anonymization on the subsequent data analysis was non-negligible and introduced quite a negative effect. Thus, summarized, to correctly interpret the effects of anonymization, as an example of important data preparation techniques, requires a lot of attention, both from a technical and domain perspective.

3.2 Modeling

The modeling phase involves the application of various techniques to build and develop predictive or descriptive models based on the selected data set. During this phase, the data mining team selects the modeling techniques that are most appropriate for the project's objectives. This can include techniques such as decision trees, neural networks, regression analysis, or clustering algorithms. The selected models are then trained using the available data, allowing them to learn patterns, relationships, and dependencies within the data. Once the models are trained, they are evaluated to assess their performance and effectiveness by the use of metrics specific to the selected methods like accuracy, precision, recall, Cohen's Kappa [7], or predictive power. If the models do not meet the desired criteria, the team may revisit the data preparation phase to improve the quality or relevance of the data or adjust the modeling techniques used. The outcome of the modeling phase is a set of

reliable, validated, and optimized models that can be used for decision-making or generating predictions.

3.3 Evaluation

The evaluation phase is crucial for ensuring the reliability, effectiveness, and suitability of the models in real-world applications. It helps in identifying any shortcomings, biases, or limitations of the models and provides insights into their overall performance. Through this evaluation, organizations can make informed decisions about whether the developed models are suitable for deployment and further utilization. Based on the evaluation results, the data mining team can make informed decisions about the models. If the models meet the desired criteria, they can proceed to the deployment phase. However, if the models fall short or do not meet the project requirements, further iterations of the modeling phase may be necessary, including refining the modeling techniques, adjusting parameters, or revisiting the data preparation phase.

3.4 Deployment

During this phase, the focus shifts from model development and evaluation to the practical implementation of the models. The data mining team works closely with relevant stakeholders and technical teams to ensure a smooth deployment process. The deployment phase aims to transform the developed models into practical solutions that can provide value to the organization or end users. It ensures that the models are effectively utilized in real-world scenarios, enabling informed decision-making, process optimization, or other desired outcomes. Any evidence of bias, unfairness, or nontransparency should be eliminated at the beginning of this phase so that the models can be put into production.

Once deployed, the models need to be continuously monitored to ensure they are performing as expected. Regular monitoring helps identify any performance degradation, data drift, or changes in the model's effectiveness. Maintenance activities may include updating the models, retraining them with new data, or addressing any issues that arise.

3.5 Data Management

The topic of data management is currently often overlooked and reduced to the part of data collection and providing the AI systems with enough (high quality) data in order to generate the best possible models. Still, data management encompasses

far more tasks, some of which directly reflect on secure long-term utilization of a model.

In order to facilitate good long-term use of data and the resulting models, *data management plans (DMPs)* [10] should be put in place, though techniques like reinforcement learning require a far more advanced approach to providing transparency than provided by the typical, rather static, approaches for DMPs. A more in-depth analysis of the shortcomings of typical templates for DMPs, including a strategy on how to overcome them, can be found in [53].

Furthermore, in order to combat the problem of hidden bias in machine learning-based systems, the DMP should incorporate information on possible sources for bias inside the data, especially when these can get propagated into the final model. This, of course, requires the owner of the system to be aware of bias inside the data.

What should not be overlooked is the importance of a proper documentation. It is crucial to document the collection and selection of training data, done in collaboration with experts, as well as the modeling and evaluation processes.

4 Security Aspects

Challenges and threats to artificial intelligence are manifold. Therefore, in this subsection, we provide an overview on the current threat landscape with regard to the machine learning system itself and in combination with cybersecurity challenges of these applications.

The threat landscape for machine learning specifically can be structured in two parts: the attack surface and adversarial capabilities. The attack surface describes where, in regards to the phase of an artificial intelligence application, an adversary might want to attack. Papernot et al. [38] define two main stages consisting of *inference* and *training*, to give a general basis, applicable to most if not all machine learning systems. Qiu et al. [39] shift these stages to be *training* and *testing*, while both describe similar things; in order to give a better overview, we propose adapting these two definitions into three phases: *inference*, *training*, and *testing*.

The inference phase includes the data ingestion, the learning algorithms, and parameters of the model and the corresponding architecture. The training phase concerns itself with running training data through the target model and building the logic and the testing phase evaluates the outputs a model produces, given a specific input.

Adversarial capabilities can be defined as the knowledge an adversary has about the artificial intelligence system and the corresponding actions they can take [38, 39]. This can be thought of rather intuitively, as access to a model directly and knowledge about its inner workings, opens a lot more attack angles than having little of either. An adversary with the former may, for example, change a model's parameters or poison ingested data during the inference and training phase.

An adversary with little knowledge and access to a target model may choose to attack a model during the testing phase, with adversarial examples [57], where

original input that was previously classified correctly, for example, an image of a stop sign being classified by an autonomous car, is tampered with in order to be classified incorrectly. These changes can be rather obvious like putting a few stickers on a stop sign, but even more robust artificial intelligence models can be lead to misclassify through the introduction of noise that can be inconceivable to a human.

The strength and severity of such adversarial capabilities are closely interlinked with an adversary's means to gain access or knowledge about a system; therefore, cybersecurity considerations become a vital component when assessing the threat landscape of artificial intelligence applications. The ENISA threat landscape report [17] analyzes cybersecurity challenges with regard to artificial intelligence, which are a comprehensive source for risk identification. In the publication, 74 threats are listed and mapped to the AI life cycle and relevant AI assets, which have been categorized into data, model, actors, processes, environment/tools, and artifacts.

In its succeeding publication [18], ENISA introduces a comprehensive guide, which analyzed more than 230 references in order to survey risk factors and their relations. Building upon a mapping between threats targeting artificial intelligence, their underlying vulnerabilities, and the AI life cycle, suitable controls have been determined, which have been categorized into the domains *organizational*, *technical*, and *machine learning specific*. For controls outside artificial intelligence, the guide references widely used standards/frameworks, such as ISO 27001 or NIST SP800-53.

A similar initiative is taken by MITRE. Following the structure of MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) matrix [37], the MITRE's ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) matrix [36] outlines threats to machine learning along the kill chain (i.e., Reconnaissance; Resource Development; Initial Access; ML Model Access; etc.). With their ATLAS Navigator toolkit tactics and techniques, MITRE provides a toolkit to link the identified techniques to threat intelligence.

The question of why an adversary might attack an AI system has to be addressed also. The number of reasons are too many to list entirely, but some examples may be the undermining of confidence, where a model could be held from deployment because a decrease in performance leads to less confidence in the models output. Another reason might be to mask certain input, for example, in network intrusion detection; a model could be meddled with to classify traffic generated by an adversary as nonintrusive. They could also go as far as damaging an organization's reputation by introducing a bias into a model that can have devastating repercussions not only for the organization in question, when becoming public, but also for people affected by the model such as credit-score estimation or artificial intelligence systems of the sort.