Rasmus Harmaala

Designing Trustworthy Algorithms

Operational Environment as a Cornerstone for Governing Risk and Enhancing Trust

School of Technology and Innovations
Master's thesis in Information Systems
Master of Science in Economics and Business Administration

Vaasa 2025

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Rasmus Harmaala
Title of the thesis: Designing Trustworthy Algorithms: Operational Environment as a Cornerstone for Governing Risk and Enhancing Trust
Degree: Master of Science in Economics and Business Administration
Discipline: Information Systems
Supervisor: Teemu Mäenpää
Year: 2025 **Pages:** 131

ABSTRACT:

Algorithms and human lives are, in many ways, symbiotic; algorithms can be found anywhere from personal devices to space exploration on Mars. We rely heavily on algorithms to assist and take care of us, but algorithms can also have fatal unintended consequences. The subject of this thesis is to find a way to increase algorithms' trustworthiness by focusing on identifying and categorizing the unintended consequences of algorithms on the operational environment of algorithms. The objective of this thesis is to provide what we call an Algorithm Power Matrix to aid algorithm developers, regulators, and private sector management in their quest for safely deploying and managing advanced algorithms on a continuous basis. Furthermore, we achieved additional progress in algorithm governance, a development that wasn't originally within the scope of this study but was enabled by our primary discoveries.

The study was structured using design science research and design science research methodology. A comprehensive literature review was conducted on algorithms and how they interact with our society. We cover the fundamentals of algorithms, how they are built, operated, and interacted with, what consequences algorithms have, and how they are regulated. This is the first phase to cover the essential parts of the study, which resulted in the creation of the first version of the Algorithm Power Matrix. The second phase involved a group of experts reviewing and suggesting editions for the second iteration cycle. The final phase was to review the iterated framework again, leading to the third iteration cycle and finalized framework presentation.

The conclusion was that even experts can have difficulties understanding how different environments affect algorithm performance. Currently, there is an intuitive way in which the developers utilize similar principles that are presented in this thesis, but this was the first time they saw something like this on paper. There is a clear need to develop more tools for understanding and controlling the effects of algorithms, and the matrix provided in this thesis is just one example. The main contribution of our research is the Algorithm Power Matrix, which is a novel framework for assessing and managing the potential impact and risk of algorithms based on their degree of freedom, autonomy, and human dependency. The matrix can help stakeholders to, most importantly, identify the appropriate level of oversight and accountability required for specific algorithms. Ultimately leading to the design and implementation of ethical and trustworthy algorithms that respect human values. Further studies are necessary to determine more precise thresholds for categorizing algorithms and their operational environment.

KEYWORDS: trustworthy and ethical algorithms, human-centered design, artificial intelligence governance, risk assessment, algorithm portfolio management

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Rasmus Harmaala
Tutkielman nimi: Designing Trustworthy Algorithms: Operational Environment as a Cornerstone for Governing Risk and Enhancing Trust
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Tietojärjestelmätiede
Työn ohjaaja: Teemu Mäenpää
Valmistumisvuosi: 2025 **Sivumäärä:** 131

TIIVISTELMÄ:

Algoritmit ja ihmisten elämät ovat monin tavoin symbioottisia. Niitä löytyy kaikkialta henkilökohtaisista laitteista avaruustutkimukseen Marsissa. Vaikka luotamme jatkuvasti algoritmien kykyyn avustaa ja huolehtia meistä, niillä voi olla myös kohtalokkaita tahattomia haittavaikutuksia. Tämän opinnäytetyön tavoitteena oli kehittää menetelmä algoritmien luotettavuuden parantamiseksi tunnistamalla ja luokittelemalla niiden toimintaympäristöjen tahattomia haittavaikutuksia. Opinnäytetyössä kehitettiin algoritmien voimamatriisi, jonka tarkoituksena on auttaa algoritmien kehittäjiä sekä julkisen ja yksityisen sektorin johtoa edistyneiden algoritmien turvallisessa ja jatkuvassa käyttöönotossa. Tutkimus tuotti lisäksi uusia edistysaskelia algoritmien hallinnassa, mikä ei ollut alkuperäinen tavoite, mutta toteutui muiden tutkimustulosten myötä.

Tutkimus toteutettiin suunnittelutieteellisellä tutkimusmenetelmällä. Tutkimuksen olennaisten osien kattamiseksi suoritimme kattavan kirjallisuuskatsauksen algoritmeista, niiden yhteiskunnallisista vaikutuksista ja sääntelystä. Kirjallisuuskatsauksen pohjalta syntyi ensimmäinen versio tutkimuksen artefaktista, algoritmi voimamatriisista. Tutkimuksen toisessa vaiheessa asiantuntijaryhmä tarkisti artefaktin ja ehdotti muutoksia sen seuraavaan iteraatioon, hyödyntäen ryhmän laajaa käytännön kokemusta algoritmeista. Tutkimuksen viimeisessä vaiheessa uudelleen iteroitu artefakti esiteltiin asiantuntijaryhmälle. Heidän palautteensa perusteella toteutettiin kolmas iteraatio ja lopullinen artefakti.

Tutkimuksen keskeinen havainto oli, että jopa asiantuntijoiden on vaikea hahmottaa erilaisten ympäristöjen vaikutusta algoritmien toimintaan. Kehittäjät käyttävät intuitiivisesti samankaltaisia työskentelyperiaatteita kuin tässä opinnäytetyössä esitetyt, mutta nyt ne on ensimmäistä kertaa esitetty kirjallisessa muodossa. Tarvitaan lisää työkaluja algoritmien vaikutusten ymmärtämiseen ja haittavaikutusten hallintaan. Tässä opinnäytetyössä esitelty matriisi on yksi näistä työkaluista. Opinnäytetyön merkittävin kontribuutio on algoritmi voimamatriisi, uudenlainen viitekehys algoritmien potentiaalisten riskien arviointiin niiden monimutkaisuuden, autonomisuuden ja vuorovaikutuksen perusteella. Matriisi auttaa sidosryhmiä määrittämään tarvittavan valvonnan, läpinäkyvyyden ja vastuun algoritmeille sekä tukee eettisten ja luotettavien algoritmien suunnittelua ja tuotantoa. Lisätutkimuksia tarvitaan erityisesti algoritmien ja niiden toimintaympäristöjen luokittelun tarkempien kynnysarvojen määrittämiseksi.

AVAINSANAT: luotettava ja eettinen algoritmi, ihmislähtöinen suunnittelu, tekoälyn hallinnointi, riskiarviointi, algoritmiportfolion hallinta

Contents

Figures

Tables

Abbreviations

Artificial Intelligence (AI)

Algorithmic Decision-Making Systems (ADMS)

Data Science (DS)

Design Science Research (DSR)

Design Science Research Methodology (DSRM)

General Data Protection Regulation (GDPR)

Intellectual Property Rights (IPR)

Machine Learning Operations (MLOps)

Machine Learning (ML)

Large Language Model (LLM)

Zero-Sum Semi-Markov Game (ZSSMG)

1 Introduction

Algorithms provide multiple benefits in different contexts in an increasing frequency, such as enabling space exploration by robots by using the Enhanced Model Reference Adaptive Control (EMRAC) algorithm as described by Montanaro et al., (2023) or as defined by Lämmerman et al., (2024) for enhancing healthcare management in clinical practice. This can be attributed to two main reasons: *reduced cost* and *increased availability of high-quality algorithms*. This has resulted in algorithms being available not just in isolation within one context, but rather that they can be implemented across numerous domains that frequently interact with one another (Martin, 2019).

The increase in algorithm usage has led to a similar rise in the unintended negative consequences of algorithm usage. This has become apparent in multiple cases arising from different contexts, such as algorithm usage in courtrooms assigning probabilities for reoccurring felonies (Rubim, 2020), faulty airplane autopilot algorithm causing two crash landings (Herkert et al., 2020), or social media used for profiling and influencing voter behavior (Atherton, 2023). As the problem is gaining more visibility in the press and academia, new technical and non-technical ways have been needed to mitigate this impact. This technological development has led to the usage of terms such as ethical or reliable algorithms, which are used to describe the new dimension of algorithm usage where there is a requirement to develop algorithms in such a way that does not cause, or at least with limits the impact, of any unintended consequences. (Mikalef et al., 2020). This increase in frequency and impact on unintended consequences is a primary motivation for this study.

As explained by Watson & Nations (2019) as well as Kordzadeh & Ghasemaghaei (2022), the increased complexity in algorithms has resulted in a demand for increased transparency and explainability. There is a long list of different attributes that guide developers to a greater understanding and explainability of their algorithms. The argument is that these guidelines when followed correctly, will contribute to the decline of unintended

consequences when developers think more thoroughly about their work. This, in turn, will lead to an increase in the trustworthiness of the algorithms. Algorithms are naturally highly trustworthy and do precisely what they are programmed to do. However, studies show numerous issues are decreasing the trustworthiness of algorithms, and the problems can be located in all parts of the algorithm development pipeline, from data to how the algorithm's final output is presented (Golovianko et al., 2022).

1.1 Objective of The Thesis

This thesis aims to increase algorithm trustworthiness by bringing together learnings from previous research and forming guidelines from the literature for the developers to increase the trustworthiness of algorithms. The foundation for these literature guidelines is to be searched from information systems literature. Our goal with this thesis is to incentivize practitioners to focus more on increasing their algorithm systems' trustworthiness. Secondly, we aim to contribute to the general awareness of regulatory authorities and the general audience by proposing a more sustainable way to implement trust in algorithms and, finally, contributing to the available toolset for designing and developing increasingly trustworthy algorithms. The first goal (incentive to increase trust) is reached using two approaches. Firstly, by providing light on the unintended consequences of algorithm systems and secondly, by introducing the increased amount of regulatory activities towards algorithms, for example, the EU General Data Protection Regulation (GDPR) and other major regulatory activities. The second goal, to provide a more sustainable and practical way to increase trust in algorithms while complying with regulatory activities, is reached by focusing on the operational environment of algorithms and by providing tools to evaluate and categorize algorithms based on their possible unintended negative impact and on the environment where they operate. The idea behind this is that there are not enough resources for extending the heaviest regulatory processes on all possible algorithms, so we should narrow the field and provide a more targeted way for regulators and companies to focus and prioritize resources on the most

critical algorithms and impose the necessary regulations and safety measures on the areas with the highest criticality.

1.2 Research Problem, Question and Approach

We can conceptualize our research motivation as the following problem statement: *increased usage of algorithms results in a similar increase in the unintended negative consequences of algorithms, decreasing their trustworthiness*. This is not a new problem, but because algorithms are widely used in different areas, new sub-problems are constantly emerging, which makes this a relatively complicated problem to grasp (Mikalef et al., 2022; Das et al., 2023). Our approach is to contribute towards the early recognition of these problems by focusing on algorithm development's design and deployment phase.

Our objectives for this thesis are the following:

1. *Examine the broad elements of algorithm development that influence the design principles.*
2. *Examine the algorithms in diverse environments and contribute towards the universal algorithm system design principles that are sector-agnostic.*
3. *Develop frameworks and tools that enable managers to effectively oversee algorithms' design, development, and deployment, ensuring they align with organizational goals and ethical standards.*

From this problem statement and these objectives, the following research questions are formulated:

RQ1: *In what ways can we identify the impacts of different algorithms?*

RQ2: *Can we measure the trustworthiness of algorithms across multiple environments?*

RQ3: *Which factors affect the scale of trustworthiness of different algorithms?*

Our research approach follows the design science research method, which is suitable for creating and evaluating artefacts that solve identified problems in a specific context. (Hevner et al., 2004; Peffers et al., 2007). The artefact we aim to develop is a framework for ethical and trustworthy algorithm design. The framework will be based on a literature review of existing theories, models, and guidelines for ethical and trustworthy algorithm design and on an empirical analysis of real-world cases of algorithm systems that have generated unintended negative consequences. The framework will be evaluated by soliciting feedback from industry experts and leaders in algorithm design and related domains. The expected contribution of this study is to provide a valuable and applicable tool for algorithm developers and managers to identify and mitigate potential unintended negative consequences of their algorithms and to enhance the ethical and trustworthy quality of their algorithm systems.

We believe that algorithms have brought and will continue to bring a multitude of advancements for humanity. Conversely, we're acutely aware of any potential societal harm as we advance in this field.

1.3 Initial Findings and Research Contributions

The trustworthiness of algorithms depends largely on their use case and the environment in which they function. By concentrating more on the characteristics of operating environments, we can classify algorithms that operate in challenging environments and determine more precisely the algorithms that require additional review and safeguards.

Our primary output from this study is an artefact called “The Algorithm Power Matrix.” This artefact serves as a framework for classifying and positioning algorithms based on their operational environment, defined by the openness and control of the operational environment. This framework aims to align regulatory requirements with private sector capabilities by identifying the algorithms with the highest probability of producing unintended consequences. Secondly, the purpose is to advise developers of the probabilities for unintended consequences on algorithms with specific properties. Our second unexpected output is the “Algorithm Portfolio Matrix,” which is a new way to utilize the Algorithm Power matrix to visualize the algorithms managed by the organization. An unexpected outcome of our study is the development of the “Algorithm Portfolio Matrix,” a novel approach to leverage the Algorithm Power Matrix in a new way to visualize, communicate, and govern the algorithms in production by an organization.

We made several contributions to the existing research, particularly in addressing the complexities of algorithmic trustworthiness in varied operational environments. By refining our understanding of these environments, we have developed tools that enhance regulatory alignment and provide actionable insights for developers, ultimately promoting safer and more reliable algorithm deployment.

1.4 Structure of The Thesis

To achieve our objectives of this thesis, we have divided the thesis into three sections. The first section is devoted to a comprehensive literature review of the algorithm ecosystem. We aim to build a solid understanding of the nature of algorithms, current design principles and workflows, production challenges, and regulatory measures. The two contemporary key concepts in this part are the differentiation of algorithm development into two parts: technology-centered review including things such as algorithm types and evaluation methods. And human-centered review including the interaction elements

between humans and algorithms including things such as artificial empathy and gaming. This section will contribute to achieving the objectives by identifying opportunities for enhancement or innovation within the design process and boosting the motivation to utilize these tools practically. Thus, it fulfils objective one and contributes to other objectives.

The thesis's second section will explain our chosen research approach, Design Science Research Methodology. We will introduce our design principles for the final artefact, build our first set of artefacts based on the literature review, and introduce it to the expert interviews. The interviews will consist mainly of experts knowledgeable in algorithmic development, complemented by business and regulatory leaders to achieve the multidisciplinary insights required for our chosen subject. We will measure our success in crafting practical artefacts based on the feedback from our esteemed experts.

Section three is devoted to the conclusion and discussion. In this part, we summarize research outcomes and present updated recommendations for practitioners while acknowledging our limitations and proposing further research directions.

2 From Code to Consequence: A Multifaceted Exploration of Algorithms

Our theoretical background begins by examining several key aspects of algorithms. We review the definition and basic structure of algorithms by introducing several diverse types of algorithms from two categories: technology-centered and human-centered. We continue by reviewing various methods of evaluating algorithms, divided into two sections: the classical evaluation and the modern evaluation methods for more sophisticated algorithms. We continue to introduce the design and workflow methods, which are key concepts for our research. Acknowledging the comprehensive nature of algorithmic studies, we add a section for a lifecycle view of algorithms. After this, we cover a few examples of unintended consequences of algorithms. We end this section with an introduction to human-algorithm interaction, examining how humans use the rules of algorithms against them to achieve different outcomes. Conversely, we provide the same service for algorithms by introducing artificial empathy.

We will finalize our literature review by providing a review of algorithm regulation. Since regulating algorithms is of paramount interest for modern economies (Rambachan et al., 2020). Our study will include the most relevant regulatory measures. Our focus group is the EU, USA, and Chinese guidelines for AI.

2.1 Understanding Algorithms: Definitions, Trends, and Challenges

This chapter will provide the basic knowledge of algorithms, including their structure, evaluation, and design.

According to the Oxford Dictionary (2024), algorithms are defined as "*a set of rules that must be followed when solving a particular problem*." Similarly, Merriam-Webster (2024)

describes it as "*a step-by-step procedure for solving a problem or accomplishing some end*." The Cambridge Dictionary's definition (2024) emphasizes the mathematical aspect of algorithms: "*a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem*." The Oxford Dictionary (2024) also notes that the etymology of the word "algorithm" traces back to ancient Greece, providing it with significant historical context. The algorithm is often found in the context of mathematics and computer science but can be applied to any field.

Sherry and Thompson (2021) indicate that the use of algorithms is rapidly expanding across various domains, leading to an increased reliance on algorithms for a vast array of tasks, ranging from organizing search engine results to developing autonomous robots for space missions. The rise in the usage of algorithms can be linked to four main trends: firstly, the reduced cost and greater accessibility of computing hardware, enabling more individuals to use algorithm-driven devices like smartphones (Botelho, 2021; Huang et al., 2023); secondly, the growing volume of data produced annually, which can be processed by these algorithms since it is available in digital format, often referred to as Big Data (Shukla et al., 2023 and Ramachandran, 2024); thirdly, technological advancements in the field of algorithms, including the emergence of artificial intelligence (AI) (Sarker, 2024 and Abdel-Karim et al., 2021); fourthly, there are more trained programmers now who integrate algorithms into their applications and the scalability of this software. Additionally, this surge of interest in computer science has spawned entirely new sectors and job roles, such as the role of the data scientist (Feuerriegel et al., 2024).

Sevilla et al., (2022) explain how these key trends have led to algorithms being employed to tackle a broader range of problems than before, including more complex ones that were not so long ago far beyond our abilities, fostering advancements in numerous domains. We're giving more decision-making responsibilities to these algorithms and increasingly relying on them for support in our decisions, facilitated by the growth in decision support systems and automated decision-making technologies. However, this heightened reliance brings challenges in comprehending the interconnectedness that

we have introduced. As the volume of available information grows and the interplay between this data and the algorithms that utilize it as input becomes more complex, it becomes harder to grasp the inner workings of these algorithms and the processes by which they generate their results.

Paradoxically, we increase the usage of algorithms to handle the increased complexities of available information. However, this creates more complex environments, which can have multiple unintended consequences, thus creating a loop of ever-increasing complexity. (Poel, 2023; Marjanovic et al., 2018).

2.1.1 Algorithm Development: Technical and Human-Centered Aspects

The study of algorithms is pervasive and varied. To give the reader a clearer understanding of recent advancements, particularly the impact algorithms have had on humans and how the interaction with algorithms has evolved vastly during recent years, we will split the examination into two parts: technical and human-centered as proposed by Wang et. al., (2024) and Xu et. al., (2021). Next, we will introduce the technical aspects of algorithm development, including algorithm structure and related workflow. Then, we will move on to algorithm development's more recent human-centered aspects.

2.1.2 The Anatomy of an Algorithm: Input, Process, and Output

Algorithms are generally viewed in the technical domain as problem solvers for computational problems, such as classification or sorting problems, and are used primarily in the domain of computer science. In its simplest terms, the algorithm can be described as having a structure of three parts. The input is the data that the algorithm uses to solve the problem, secondly the method for reaching the desired output, which is the algorithm itself, which in itself is the set of mathematical rules determining how the algorithm is intended to solve the given problem and finally the output which is the result

and should present a solution to any given problem. As an example, the input might be a list of webpages, the algorithm a sorting algorithm, and the output would be a re-arranged list of these webpages to match the defined criteria, such as which one is the best match for finding recipes (Cormen et al., 2009 pp. 5-7).

2.1.3 From Design to Deployment: Algorithm Development Methodologies

To design and create trustworthy and ethical algorithms, it is essential to understand the common approaches involved in their design and implementation (Saltz & Dewar, 2019). Algorithms are typically designed to serve a specific purpose, such as categorizing objects according to their characteristics or optimizing routes for navigational purposes. While we can sometimes recognize how and where algorithms function, there are instances where their work goes largely unnoticed, like in automated loan processing or hyper-personalized marketing, which can be seamlessly integrated into products that go into our daily activities without us having transparency on their design principles (Dourish, 2016; Krishnaraj et al., 2024; Tong et al., 2020). Understanding the methodologies that software teams employ to develop applications using algorithms is a significant focus of our research. To gain an understanding of this area, we will introduce two models for algorithm development. The Data Science Trajectory (DST), which is an updated version of the classical and commonly used Cross-Industry Standard Process for Data Mining (CRISP-DM) model, which, although developed in the 1990s, remains the de facto framework for executing data science and algorithm-driven projects (Saltz & Krasteva, 2022). Additionally, we will present the Machine Learning Operations (MLPOps) model, a more recent approach in algorithm design primarily aimed at managing Machine Learning (ML) projects.

Martinez-Plumet et al., (2021) have introduced a DST workflow, which represents a modern approach to the data science process, updating the traditional CRISP-DM. This model has been predominant since 1999 (Dastgerdi & Gandomani, 2021). The principal distinction between the original CRISP-DM, which consisted of 6-stages and represents

significant similarities with the updated model, and today's practice lies in the term "mining." In 1999, the data science sector was still in its infancy, primarily focused on rule-based algorithms, and leveraging data science tools only for well-defined objectives. Fast forward to the present, we have access to vast quantities of structured data, allowing us to use data science not just for targeted analysis but for exploration, uncovering insights we might not have initially expected. The approach has evolved during the past decade or so and while our end goals may not always be clear from the beginning, we have discovered sophisticated algorithms that help us derive value from data in new and unexpected ways (Martinez-Plumed et al., 2021).

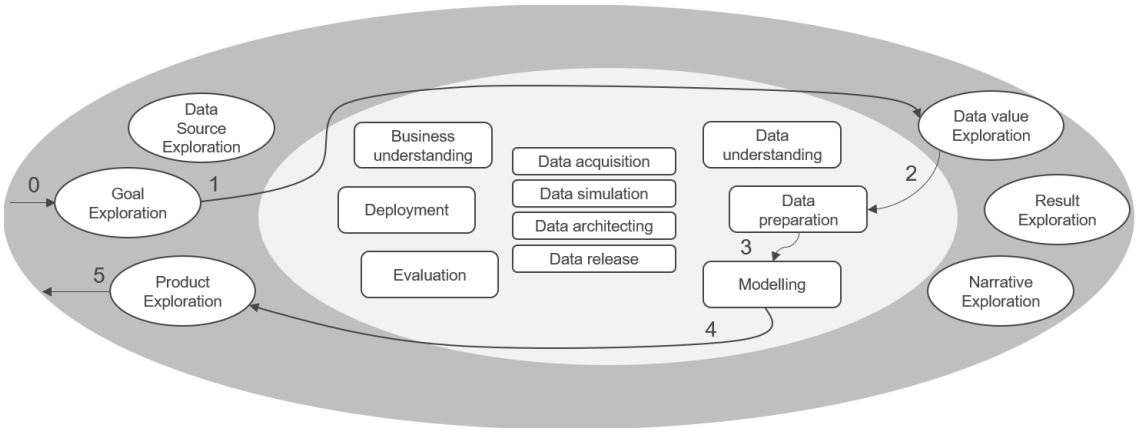

Figure 1. Example trajectory of a DS project (Martinez-Plumed et al., 2021).

As we can see from the data science trajectory, considerable space is dedicated to data handling. As Ntoutsi et al., (2020) and Jaigirdar et al., (2020) point out, properly handling data is essential for achieving a trustworthy algorithm. Any error in the data set can lead to undesired consequences such as overfitting, biased decision-making, or unexplainable behavior. An example of data handling gone wrong is the Danish Healthcare data collection project DAMD, as explained by Aaen et al., (2022), where good intentions of having a quality database for treating patients became a privacy nightmare and resulted in the ending of over 10 years of data collection project. It is, therefore, crucial to thoroughly understand your data sources, safeguard them for privacy, and maintain suitable access control.

Early models such as CRISP-DM have faced criticism for not adequately involving multiple stakeholders and failing to clearly define required team member roles and objectives, which are essential in modern data-intensive projects (Saltz, 2021; Li et al., 2023). According to Eken et al., (2024), the concept of MLOps was created to address these new challenges and provide a model for developers to integrate and orchestrate all the necessary components in ML projects. Kreuzberger et al., (2023) note that many ML projects do not meet management expectations and fail to add value, often leading to abandoning the project. This perspective highlights our research that even if you develop a successful ML project with cutting-edge algorithms, it can fail if an unintended negative impact occurs. Therefore, it's crucial to consider the entire workflow from design to lifecycle management, which involves multiple implications for various parties involved, especially for the end-users throughout the project. To highlight the differences between these two approaches, DST and MLOps, we will add the entire MLOps pipeline and workflow adopted from Eken et al., (2024) for the reader's comparison.

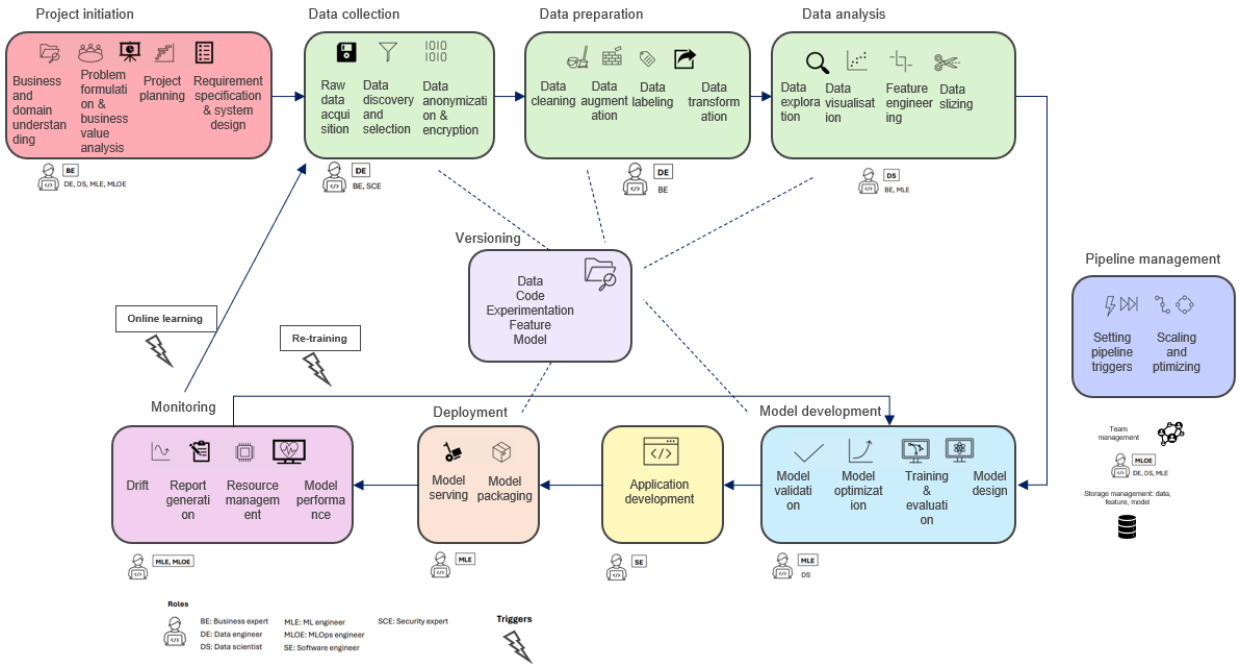

Figure 2. A pipeline for ML development and operations (Eken et al., 2024).

Figure 2, the ML development pipeline provides a clear picture of the ML development project workflow. The categorization of each activity is more expanded than the model from Figure 1. As we can see from Figure 2, there are more activities defined than in Figure 1. Also, figure 1 lacks any role descriptions compared to Figure 2, where roles have been dedicated to the related activities. Figure 2 emphasizes two aspects of algorithm development not mentioned in Figure 1: iteration and monitoring. Figure 2 connects iteration with every part of the development and adds the monitoring to the end of the project.

2.1.4 Navigating the Complexities of Algorithm Development: Lifecycle and Governance Considerations

Understanding the reliability of algorithms requires recognizing that algorithm development is a complex, multi-stage process, with multiple steps potentially impacting their trustworthiness. There is a constant need to balance the algorithm between performance and its trustworthiness: enhancing trustworthiness typically requires more time for implementation and at the expense of the algorithm's performance or accuracy (Nguyen et al., 2021). To deepen our understanding of the characteristics and interactions that influence algorithm development, we will present the reader with two models: firstly, the Algorithm Decision Making Systems (ADMS) Lifecycle model by Marabelli et al., (2021), and secondly, the AI hourglass governance framework from Mäntymäki et al., (2021). The ADMS model is a more detail-oriented, while the hourglass governance framework offers a broader organizational perspective. Both are vital for the reader to comprehend algorithm development's entire lifecycle and impacts.

We will first define key terms for this chapter. Firstly**,** *AI assurance* refers to the auditing of algorithms to ensure they are safe, reliable, and trustworthy. This is done in four key categories, as Freeman et al., (2022) and Batarseh (2021) explain. Firstly, Verification & Validation refers to the error-free system. Secondly, Development & Deployment refers to the unique demand of the environment where the algorithm is deployed and the

specific testing for that system. Thirdly, Transparency: algorithm output should be understandable and easy to interpret. Fourthly, bias usually refers to the over- and underfitting of the model. Finally, Context: this can also be called “situational awareness” and refers to the unique setup and demands of the environment where the algorithm operates and its unique challenges.

AI Governance, as described by Sharma (2023), Mäntymäki et al., (2023), and Batool et al., (2023), is a partial solution for a complex problem involving the incorporation of technical, legal, and organizational aspects such as processes and frameworks for safe, ethical, and effective use of AI systems. AI governance aims to align all the layers related to AI as a guiding principle for the organization to use when dealing with AI systems.

Marabelli et al., (2021) have identified key strategic choices regarding this balance that developers should consider during the algorithms' lifecycle. These choices are presented in the following figure.

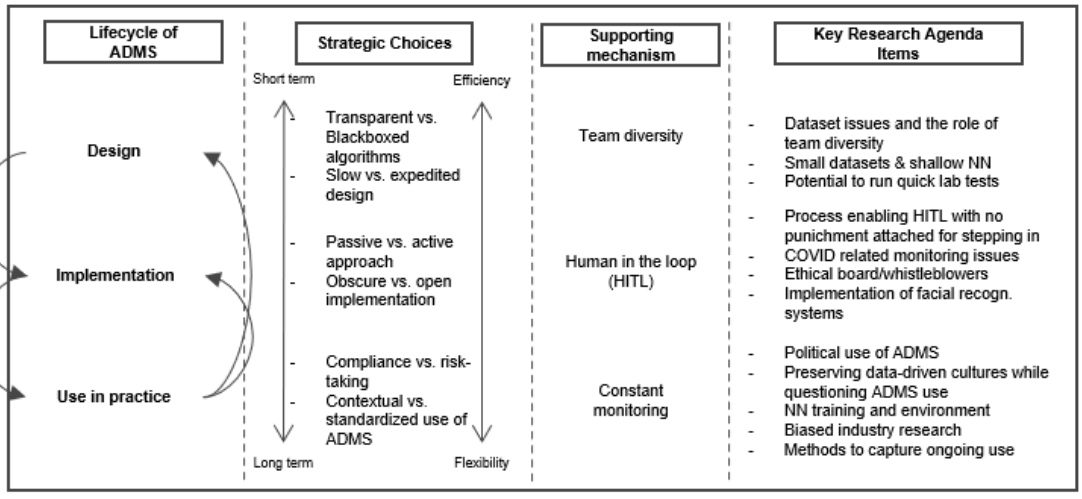

Figure 3. Managing strategic choices in the lifecycle of ADMS (Marabelli et al., 2021).

Marabelli et al., (2021) emphasize that the process of developing algorithms is traditionally headed by technical people such as ML engineers, who tend to prioritize technical features and artefact’s ability to reach the desired output, sometimes overlooking the

human aspect. This oversight can lead to unwanted negative outcomes. Therefore, there should be an advocate for a collaborative approach to design and development, involving technical experts alongside human science specialists, such as anthropologists or psychologists, to foresee better and integrate the impact on people. Reducing or minimizing unintended negative consequences should be a priority.

AI governance is a growing topic and considers the aspects that the developers might not consider; often, these are the external factors that might impact the algorithm development (Batool et al., 2024; Danaher et al., 2017). The development of governance frameworks has increased, and there are now several frameworks for consideration, such as the hourglass model, NIST AI Risk Management Framework, OECD Principles on Artificial Intelligence, IEEE Global Initiative on Ethics of Autonomous *and Intelligent Systems, Montreal Declaration for Responsible AI* and *the GAO AI Accountability Framework*. The governance model aims to establish a framework for ensuring that AI development is thoughtful and in line with organizations and legal requirements. We will now introduce one of these AI governance frameworks to give the reader an overview of the topic, the hourglass model by Mäntymäki et al., (2023).

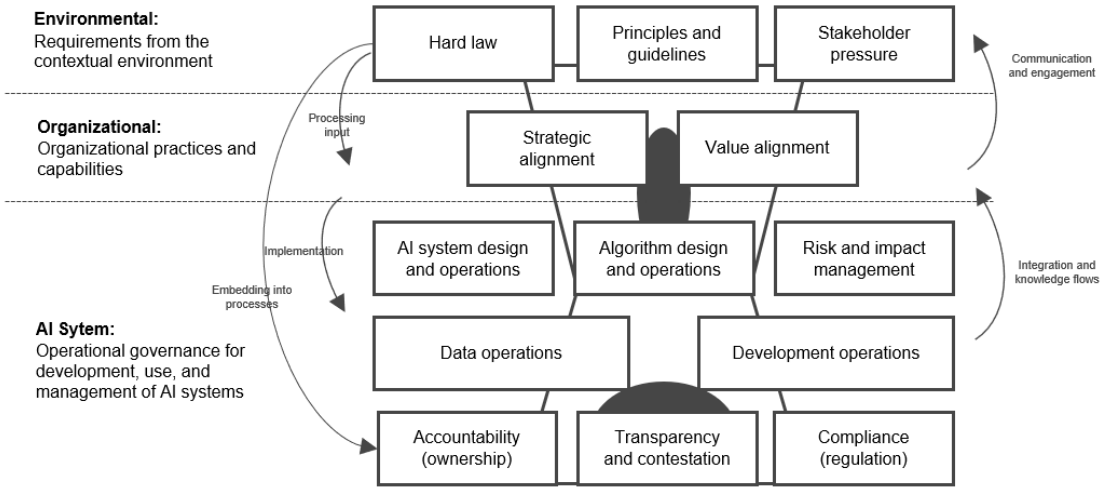

Figure 4. The hourglass model of AI governance (Mäntymäki et al., 2023).

The hourglass model comprises three distinct layers, as illustrated on the left side of Figure 4. The first is the environmental layer, which addresses external relations that influence and require reporting from the organization. The second, the organizational layer, deals with internal alignment within the organization, ensuring proficient guidance for algorithm development and alignment with other internal activities. Lastly, the AI system layer focuses on the detailed construction of the algorithm systems. This layer also initiates the reversal of the information flow, sending data back to the top of the hourglass to keep the upper layers informed (Mäntymäki et al., 2023).

2.1.5 A Taxonomy of Algorithms: Classical, Machine Learning, and Beyond

Next, we will introduce a few different types of algorithms. Categorizing algorithms is not an easy and unambiguous task; algorithms can be categorized, for example, by a*) by the method it solves the given problem* or b) *by the problem it is intended to solve* (Ezugwu et al., 2024). As explained by Zhang & Lvzhou (2022), the discoveries in quantum computing have sparked a categorization where all algorithm types are divided into one of the following: classical, quantum, or hybrid algorithms. The main difference between classical and quantum algorithms, as explained by Hughes et al., (2021, p. 81-84), is a concept called parallelism, which refers to the qubits in quantum computers having simultaneously multiple parallel positions called superposition. The quantum aspect is above the scope of our study, but it provides a useful comparison for categorizing algorithms based on the computer's hardware. Classical algorithms run on classical computers, quantum algorithms run on quantum computers, and hybrid algorithms run on hybrid computers, usually a supercomputer, which is integrated with a quantum computer. In the interest of clarity, we will deviate from this categorisation slightly by grouping algorithms into classical and ML. Next, we will introduce a few classical algorithms and algorithms which are capable of learning. We will continue introducing additional algorithms in the next chapter that focus more on human-centric design principles.

There are numerous algorithms for solving countless different kinds of problems. The first group we will examine is classical algorithms, considered the earliest and most robust algorithms. These algorithms are used as a baseline for enabling more advanced algorithms to operate at their full capacity. Bullynck (2015) and Tindall et al., (2024) introduce four classical algorithms that enable clean and useful data infrastructure, which is a must for data-intensive requirements of ML applications. The most essential classical algorithms used in computer science are sorting, searching, graphs, and string.

Sorting refers to rearranging data to a desired order of importance. All computer systems utilize sorting algorithms in different ways, such as arranging web pages. Sorting algorithms increase efficiency when used with other algorithms and improve their performance in today's world, where we have trillions of data points in massive data centers. It would take considerably more time to find the exact data you need unless you have a way to sort the available data (Pizarro-Vasquez et al., 2020). **Searching** is a way to find relevant data you're looking for, such as the minimum and maximum values in any given data set. This is also an elementary part of enabling the usage of advanced algorithms (Xing & Marwala, 2017). **Graphs** represent the relationships and connections between different data points. Graph algorithms are used, for example, to determine the fastest route between two points and the strength of the connectivity between the two points. For example, how strong a relationship different categories of books have in a bookstore (Wills & Meyer, 2020). **String** algorithms transform any given data into a more easily readable digital language that computers can process faster (Ferragina, P., 2008).

We can extend the types of algorithms to include more advanced algorithms that constantly iterate their processes and seek to learn from the input data. This group is referred to as Machine Learning (ML) and Artificial Intelligence (AI) algorithms.

ML algorithms refer to a concept used in computer science where the algorithm independently utilizes its own past experiences to learn and improve from its previous state. This algorithm style is instrumental when there is an extensive data set to be handled

and when it is tough to figure out adequate solutions by traditional algorithms (Brink et al., 2017: 3. Alpaydin 2010: 1–3). ML algorithms are especially good at discovering insights from data that humans find hard to recognize since they use advanced mathematical models that can be layered in multiple different layers, creating a complex structure that is hard to understand. The problem with ML algorithms is the data, which can be corrupted, noisy, or biased. The data can also be intentionally manipulated to misguide the algorithms utilizing this data (Alpaydin 2010: 30–31. Zhang et al., 2014). This is one aspect that is critical when designing trustworthy algorithms. Next, we will examine a few of the most common ML algorithms.

We will first examine **supervised learning**. It is a method where the algorithm utilizes pre-determined examples where the input and output are known and thus can be used in managing the algorithm's performance. This style of ML is not just helpful but also highly adaptable. It can be used in a variety of situations where you need to utilize similar solutions to known problems, such as classification problems, which are problems where you need to determine the class where a particular variable belongs correctly (Alpaydin 2010: 5–8; Brink et al., 2017: 8).

Next, we will examine **unsupervised learning**. In contrast to supervised learning, where the output is known and can be utilized to oversee the algorithm's performance, in unsupervised learning, the output is unknown and can’t be used to supervise the algorithm. This style of ML is especially thrilling when it comes to pattern recognition. Unsupervised learning is the key to discovering hidden patterns and structures in data, creating data clusters based on the data’s density estimation (Alpaydin 2010: 11).

Other common ML algorithms are **reinforcement learning**, which according to Shaza et al., (2025) is a similar approach for the way humans and animals learn by observing the environment and receiving cues and reward or penalty signals for their actions. **Deep reinforcement learning** is a very similar approach to reinforcement learning but can be used on larger problems since it has the capacity to generalize the possible outcomes

which reduces the required resources and allows it to be used on larger problems (Gabriel & Madeira, 2024). According to Reijonen (2018), other common algorithms are **dynamic learning**, which refers to real-time learning and **transfer learning**, which refers to a way to use same algorithm on different problem set from the original. There are many different learning models, and the ML field is constantly developing and creating new learning methods to increase performance and solve new kinds of problems (Reijonen, 2018).

Below, we have provided a summary table (Table 1) of the previous algorithms for the reader's convenience.

Table 1. Summary of Classical and Machine Learning Algorithms.

Algorithm Type	Algorithm	Summary	Sources
Classical	Sorting	Rearranges data into a desired order, improving efficiency and performance of other algorithms.	Pizarro-Vasquez et al., (2020)
Classical	Searching	Finds relevant data within a dataset, such as minimum and maximum values.	Xing & Marwala (2017)
Classical	Graphs	Represents relationships between data points, used for tasks like finding the fastest route.	Wills & Meyer (2020)
Classical	String	Transforms data into a readable digital format for faster processing by computers.	Ferragina (2008)
Machine Learning	Supervised Learning	Uses pre-determined examples with known input and output to manage algorithm performance.	Alpaydin (2010); Brink et al., (2017)
Machine Learning	Unsupervised Learning	Discovers hidden patterns in data without known output, creating data clusters.	Alpaydin (2010)
Machine Learning	Reinforcement Learning	Learns by receiving rewards or penalties for actions, optimizing behavior over time.	Shaza et al., (2025)
Machine Learning	Deep Reinforcement Learning	Similar to reinforcement learning but can handle larger data sets by generalizing possible outcomes.	Gabriel & Madeira (2024)
Machine Learning	Dynamic Learning	Continuously adapts and learns from new data in real-time.	Reijonen (2018)
Machine Learning	Transfer Learning	Applies knowledge gained from one problem to a different but related problem.	Reijonen (2018)

2.1.6 The Human Factor in Algorithms: Exploring Explainable, Trustworthy, and Generative AI

As algorithms have advanced tremendously in recent years, as computing power and data have increased, the human element has begun to attract increased attention, as Minh et al., (2022) explained. Advanced algorithms have now reached more users than ever, resulting in the need to increase the examination of human-algorithm interactions. One notable example of this trend is the founding of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). Founded in 2019 to ensure that AI development prioritizes human well-being (Stanford Institute, 2024).

Next, we will examine a couple of new definitions related to algorithms, mainly on the more advanced algorithms such as Explainable and Interpretable AI; we will also examine trustworthy AI. These definitions have been introduced after the realization that there are unintended biases and consequences on algorithms, and these definitions try to set up a foundation to address these flaws in algorithms and advance a more holistic approach to the usage of algorithms. They have a few common aspects, and most notably, they are designed to be more human-centered, which means including the human element in the design phase. We will finish the chapter with an overview of generative AI and new techniques related to it.

Next, we will examine **Explainable** and **Interpretable AI (XAI** and **IAI)**. One could believe that the recent advances in AI have raised the need to understand the reasoning behind AI decisions, but this is not the case. The need for explainable AI is as old as AI itself, and researchers have always had the need to understand their models in detail; only recently has this become the interest of a more general audience (Meske et al., 2022). As with most terms related to AI, there is also a need for clarification on the definition of explainable and interpretable AI, which are often used as synonyms. Interpretable AI is defined as a more passive way to understand AI without a more detailed active explanation. XAI is understood as a requirement for deeper understanding and explainability, such as in

the case where the AI logic is not clear, for example, in the case of Deep Learning models (Guidotti et al., 2018; Abadi et al., 2018; Meske et al., 2022).

Creating more explainable AI is a wide-reaching goal. It has been recognized as a goal in many organizations, such as in the EU AI Strategy and, for example, in the Defense Advance Research Projects Agency (DARPA), which is the US Military Research Centre for Advanced Technology. In 2017, DARPA launched a specific program to develop more explainable AI; here is how DARPA defines the XAI:

> *"DARPA defines explainable AI as AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. Naming this program explainable AI (rather than interpretable, comprehensible, or transparent AI, for example) reflects DARPA's objective to create more human-understandable AI systems through the use of effective explanations. It also reflects the XAI team's interest in the human psychology of explanation, which draws on the vast body of research and expertise in the social sciences."* (Gunning & Aha, 2019, p. 44).

Meske et al., (2022) also point out in their research (citing Abdul et al., 2018; Fernandez et al., 2019; Miller, 2019) that explainability should be considered a prerequisite for fair, accountable, and trustworthy AI and the absence of explainability would eventually lead to the increased probability of undesired consequences since there is no way to perform reliable threat analysis on non-explainable AI models, resulting ultimately in uncontrollable AI. This is one of the key findings and a source of inspiration when we present our "AI Power Matrix" and "Operational Environment" considerations later in the study.

Next, we will examine the **Human-Centered AI (HCAI)**. The HCAI approach emphasizes the AI design, where the human aspect is considered in every phase of AI development. The goal is to enhance machine-human performance by creating reliable, trustworthy, and safe AI systems. Usually, this consists of two approaches: firstly, the explainability of the AI towards humans should be increased, and secondly, the explainability of humans as a socio-technical environment where algorithms are just one part should be

explained to the AI (Rield, 2019). HCAI has achieved traction, especially in the Human-Computer Interaction (HCI) community, where the emphasis is placed on the user experience (UX) and the interface design (Schneiderman, 2020).

Next, we will examine **Trustworthy AI**. Trustworthiness refers, in general, to a way where one has confidence that someone will complete the requirements set to her and achieve agreed objectives. Arrieta et al., (2020) describe trustworthy AI in a similar manner and add that trustworthiness is one of the essential requirements and most important aspects of explainable AI. Trustworthiness in Information Science is usually used to refer to the quality of information, systems, or entities (Cho et al., 2019). Arrieta et al., (2020) pointed out some conflicts in the usage of trustworthy by pointing out that ethical AI might not be trustworthy, that trustworthy AI might not be ethical, and that ethical AI should be chosen over trustworthy AI.

Both of these definitions have their place since, as Raden (2020) pointed out, they are not explicit, and one can exist without the other. Thus, there is a need to identify algorithms as trustworthy, meaning they will accomplish their objective, and ethical, meaning they will achieve the given objective with high moral standards and without causing harm to others.

According to Arrieta et al., (2020), trustworthiness is one of the core elements in modern AI systems. However, this is not directly similar to explainable AI since trustworthy AI doesn't necessarily require the ability to explain its actions. There is a key difference here. Trustworthy algorithms can be trusted to do the job and achieve the required output in a way you desire, but not necessarily with 100% transparency. (Arrieta & alt, 2020; Ribeiro, Singh, Guestrin, 2016; Shneiderman, 2020).

Trustworthy algorithms should act as intended by their developers. Therefore, any algorithmic system producing unintended consequences is, by definition, untrustworthy (Arrieta et al., 2020). When algorithmic systems became increasingly more complex,

achieving a system without any deviations from the intended state was more challenging. As mentioned earlier, issues may occur in numerous areas in the development phases, and achieving zero unintended consequences is very hard. This is also one reason why we will bring increased focus to the operational environment where the algorithms will eventually be deployed.

Even if algorithms don't achieve complete trustworthiness, they are not necessarily immediately categorized as distrustful. According to Benamati et al., (2010), distrust arises when there is an explicit action towards incompetency, neglectfulness, or harmfulness. In the case of algorithms, an algorithm capable of learning itself might be able to reach a confidence level that is causing some of these features of distrust to the users and modify its behavior accordingly.

Next, we will examine **Generative AI (GENAI)**. According to Kankanhalli (2024) and Nah et al., (2023), Generative AI refers to a set of AI models that possess a generative ability to react to a broad set of prompts or tasks from the user. This can be used to create novel data such as text, images, or videos by interpreting and manipulating pre-existed data. Generative AI models don't create new data but rely on the historical data they have been trained. The pre-training phase of generative AI involves Foundational models, which are large AI models involving, in many cases, billions of data points from a broad range of different topics. Foundational models take time and resources to be fully trained and are used as a base layer for Generative AI models. According to Bommasani et al., (2021), Foundational models are based on Deep Neural Networks and Self-Supervised Learning, which are already decades-old technologies and only now possible to be utilized fully since they require vast amounts of data and computing resources. The new thing about Foundational Models is the scale and scope of the data in which they are trained; never before has this vast amount of data been used to train AI models. This scale creates unique phenomena such as the *AI Hallucination*, which, according to Maleki et al., (2024), refers to the cases where Large Language Models (LLM) confuse different domains and mix them together, creating a fictional answer that sounds true.

Another issue, according to Bommasani et al., (2021), is *homogenization*, which refers to the underlying similarity of data used in the AI model training. Foundational pre-trained models are usually heavily re-used, which means they inherit the *Data Bias* and other possible underlying issues, which transfer then to the end model.

We will briefly cover a set of new techniques developed especially for LLMs to increase their accuracy and reasoning capabilities, as explained by Wei et al., (2022). *Chain-of-Thought* (CoT), aims to systematically break down the problem presented for the LLM and then individually answer each part of the question, as explained by Wei et al., (2022). Wang et al., (2022) introduced a *Self-Consistency* (SC) reasoning which aims to increase the answer's consistency by always choosing the most consistent path from multiple options. Yao et al., (2023) proposed a *Tree-of-Thought* (ToT) reasoning technique that explores a multitude of reasoning paths and forms a tree-like structure before choosing and presenting an answer based on the most optimal path. Chu et al., (2023) introduced an improved technique based on the ToT, *Graph-of-Thought* (GoT) which aims to strengthen the connections between interrelated concepts allowing the model to enhance confidence in a multi-variable environment. *Program-of-Thought* (PoT) presented by Chen et al., (2023) allows the LLM to structure the question in a pseudo-code, thus increasing the understanding of the question. *Algorithm Distillation* (AD) introduced by Hu et al., (2024) is a method to train the LLM algorithm to incorporate any of these systematic reasoning techniques from the very early training phase instead of adding these techniques after the training phase.

Below, we have provided a summary table (Table 2) of the previous algorithms for the reader's convenience.

Table 2. Summary of Human-Centered Algorithms.

Algorithm/Concept	Summary	Sources
Explainable AI (XAI)	AI systems that provide understandable explanations for their decisions and actions.	Meske et al., (2022) Guidotti et al., (2018) Abadi et al., (2018) Gunning & Aha (2019)
Interpretable AI (IAI)	AI systems that allow users to understand their decisions without detailed explanations.	Meske et al., (2022) Guidotti et al., (2018) Abadi et al., (2018)
Human-Centered AI (HCAI)	AI design approach that considers human aspects in every phase of development.	Rield (2019) Schneiderman (2020)
Trustworthy AI	AI systems that are reliable, safe, and meet ethical standards.	Arrieta et al., (2020) Cho et al., (2019) Ribeiro, Singh, Guestrin (2016) Shneiderman (2020)
Generative AI (GENAI)	AI models that create novel data such as text, images, or videos based on pre-existing data.	Kankanhalli (2024) Nah et al., (2023) Bommasani et al., (2021) Maleki et al., (2024)

2.1.7 Evaluating Algorithms: Classical and Modern Analysis

This section will be divided into two parts: classical and modern evaluations. The classical part refers to the early algorithms, and the contemporary part covers the more advanced ML algorithms. According to Cormen et al., (2019) and Handelman et al., (2019), there are two main targets for evaluating algorithms. The first is to monitor the algorithm's performance, such as the ratio between true and false positives or negatives and how much variability there is between the predictions. The second target is to monitor the resources the algorithm utilizes to achieve its performance. We classify these as software and hardware evaluations.

According to Cormen et al., (2009), the **classical way** to evaluate algorithms is by technical performance in *time* and *space* categories, meaning how long it takes the algorithm to reach its objective (time) and how much computational resources such as memory and processing power (space) it takes. This is called the analysis of algorithms, and it is used to determine if the computational process is efficient in solving the given problem at hand. Classical algorithms can only sort a problem of a certain level of difficulty in a

reasonable time and with reasonable resources. As we face ever-growing amounts of more complex problems, we wish to solve so, too, that we have the need to build more efficient tools to solve them. In the world of algorithms, this means the transformation from classical to modern and, eventually, quantum algorithms. Beiranvand et al., (2017) suggest that we use ML algorithms because they are more efficient, meaning they take fewer resources (space) and they solve a problem faster (time) than classical algorithms. We don't have the time and resources to wait for an algorithm to solve a problem in a thousand years if we can solve it in an hour by employing more advanced algorithms. This is why we analyze algorithms to determine if they can solve the problem in a reasonable time and space. Since speed is key with algorithms, faster and more efficient algorithms offer a more pleasant customer experience and provide a competitive advantage to the companies.

With the scalability effect in algorithm development, even a 1% efficiency increase is considered a good achievement and should be taken seriously in algorithm evaluation (Cormen et al., 2009). Beiranvand et al., (2017) recognize that achieving similar improvements is a tedious task and propose adding automation to optimize the algorithm performance based on several evaluation metrics.

Next, we will consider **modern algorithm evaluation**. According to Rainio et al., (2024), The contemporary evaluation of algorithms has made significant developments during the last decade, such as focusing on developing robust statistical testing for specific ML algorithms and increasing interdisciplinary collaboration. Modern algorithm evaluation has developed two distinctive targets. Firstly, the performance of the new and old algorithms is compared to determine which one is superior, and secondly, the superior algorithm is optimized further. The evaluation metrics have also developed from more statistical versions to more task-specific evaluations. This has been driven by the advancement of advanced algorithms, which are in many ways more complicated than the classical algorithms. Space and time as evaluation metrics are still relevant, but there is a need for more accurate measures. This is mainly because the number of different

algorithms has increased, and their differences don’t allow them to be sufficiently evaluated by using the same metrics with slightly different algorithms. Bandi et al., (2023) propose that modern advanced algorithms have also been developed to be more oriented toward specific tasks, such as language or image processing, and it's more accurate to measure the algorithms designed to perform similar tasks.

Evaluation metrics for advanced algorithms are primarily focused on whether the algorithm provides the correct answer and, if yes, how many times and with what kind of deviations. The questions should also be domain-specific. We will briefly introduce a few key metrics to evaluate advanced algorithms. Naidu et al., (2023) present several evaluation metrics for advanced algorithms, such as *Accuracy* and *Precision*, accuracy refers to the ratio of correct and incorrect predictions. Precision examines the ratio between true positives and total positives. Naidu et al., (2023) also present ROC and AUC, *Receiver Operating Characteristics*, and *Area Under the Curve* are used to measure the performance of a classifier across different thresholds. ROC is the ratio between true positive rate and the false positive rate. AUC refers to the algorithms probability for correctly predicting these true positives and false positives. Rainio et al., (2024) group these metrics as Binary Classifications and introduce five other groupings of key evaluation metrics that mainly utilize statistical methods in evaluating the performance of algorithms.

The evaluation metric for GenAI has been expanded to include the success rate of bypassing safeguards, as described by Phuong et al., (2024) and Wei et al., (2023). Interest in LLMs is due to their numerous capabilities, which can be misused if proper safeguards are not in place, as Bubeck et al., (2023) explained. According to Hughes et al., (2024), one way to test and evaluate LLM safeguards is the Best-of-N (BoN) jailbreaking algorithm, which generates automated prompts with slight variations to challenge the safeguard. Researchers have achieved up to a 78% *Attack Success Rate* (ASR) using the BoN algorithm on several known LLMs (Hughes et al., 2024). Another recent development in the LLM field is the increase in sector-specific safety and evaluation tools, such *as CYBERSECEVAL*, designed to review the code provided by LLMs from a cybersecurity

standpoint, as explained by Bhatt et al., (2023). Another standard method to evaluate GenAI algorithms, due to their broad applications, is *Red Teaming*, where developers proactively try to find ways to exploit the system and elicit unintended responses (Kosinski et al., 2023).

Below, we have provided a summary table (Table 3) of the previous algorithm evaluation metrics for the reader's convenience.

Table 3. Summary of Algorithm Evaluation Metrics.

Evaluation Type	Metric	Summary	Sources
Classical	Time	Measures how long it takes for an algorithm to reach its objective.	Cormen et al., (2009)
Classical	Space	Measures the computational resources (memory, processing power) required by an algorithm.	Cormen et al., (2009)
Modern	Accuracy	Ratio of correct predictions to total predictions.	Naidu et al., (2023)
Modern	Precision	Ratio of true positives to total positives.	Naidu et al., (2023)
Modern	ROC (Receiver Operating Characteristic)	Measures the ratio of true and false positives across different thresholds.	Naidu et al., (2023)
Modern	AUC (Area Under the Curve)	Measures the overall performance of classifying either false or true positive.	Naidu et al., (2023)
Modern	Attack Success Rate (ASR)	Measures the success rate of bypassing safeguards in generative AI models.	Hughes et al., (2024)
Modern	CYBERSECEVAL	Reviews the code provided by LLMs from a cybersecurity standpoint.	Bhatt et al., (2023)
Modern	Red Teaming	Proactively finding ways to exploit the system and elicit unintended responses.	Kosinski et al., (2023)

2.1.8 The Complexities of Human-Algorithm Interaction: Gaming, Empathy, and Collaboration

In this chapter, we will define human-algorithm interactions and then examine three interconnected key concepts: ***gaming algorithms***, ***artificial empathy***, and ***game theory***. Our primary focus will be on the ways humans and algorithms try to deceive each other for personal gain, which will be covered in the first two key concepts; the final key

concept is related to finding common ground between humans and algorithms for mutual benefit and emphasizing cooperation instead of personal benefit. Human-algorithm interaction, also known as human-AI interaction (HAX), involves how humans and algorithms engage through any accessible interface, such as an application, computer screen, or even voice calls. According to Tarafdar (2023), humans and algorithms typically interact via dedicated apps such as Uber, a ridesharing app. Uber is an example of HAX, where algorithms can evaluate and even assign tasks to humans during interactions, creating time ambiguity and uncertainty for humans. The crucial element in HAX is having a common language and mutual understanding. However, this can be challenging since algorithms, which rely on complex computations, can be difficult for humans to comprehend. Likewise, algorithms may struggle to grasp human communication as our language continuously evolves (Page et al., 2017). Tsiakas & Murray-Rust (2024) highlight two distinctive challenges within the HAX space, including *capability uncertainty*, meaning the unpredictability of understanding what algorithms are capable of, and *output complexity*, which refers to the difficulty in comprehending the implications of multidimensional, intricate outputs from extensive algorithmic systems. The latter is particularly relevant to our research, especially in the forthcoming chapter.

Much like their developers' algorithms are not perfect. Instead, they are prone to the same vulnerabilities and values to which they have been programmed to comply, and users understand these limitations (Burrell, 2016; Eubanks, 2018). This imperfection provides an opportunity for the users to reach the desired outcome, which might not be intended to be possible by the developers, by manipulating the algorithm, for example, using *AI prompt engineering*. Users engage with these algorithms in an attempt to decipher their governing rules and exploit this knowledge to manipulate the system. Such actions may secure them a better standing, like advancing to a job interview with a recruiter or achieving a higher rating in an insurance company's assessment, which could result in reduced insurance premiums. This behavior is commonly referred to as "*algorithm gaming*," as Petre et al., (2019) and Ziewitz (2019) explained. Next, we will examine this concept in depth. **Gaming algorithms,** defined as a "*purposeful change in order to*

alter the algorithm's estimates" (Bambauer & Zarsky, 2018, p. 10). As explained by Petre et al., (2019) and Zarsky (2016) gaming can be an unharmful activity performed by the user's intention to test the model's limits or behavior for unusual input out of curiosity or fun. It can also be performed with malicious intent to make the algorithm perform actions that are not the original intent or have been red-flagged by the developers, such as providing instructions to create homemade explosives or impersonating a company director and using AI-generated fake voices to steal 35 million dollars (US DOJ, 2021). Gaming algorithms require a deep understanding of the limitations and functionalities of the relevant algorithms, and equally, the prevention of this phenomenon requires considerable effort and knowledge of the algorithms from their developers.

A notable example of gaming would be the Microsoft chatbot Tay in 2016, designed to conduct conversations with a wide-ranging public audience on various topics. Once it was released in public, the general audience quickly realized how to "*game*" the bot, and Tay was coerced to adopt Nazism ideologies and declared at one point that "*Hitler was right*" and other racist comments, for example, targeting feminist sympathizers. Microsoft had to retire Tay soon after, even though Microsoft had a novel goal with Tay, it turned out to be a disaster after users learned how to game the algorithm behind it (Schwartz, 2019).

Another case in point is the British artist James Bridle, who conceptualized "*Autonomous Trap 001*," an artwork in which he created a "trap" for an autonomous vehicle using salt to draw a circle of "Do Not Cross" or "No Entry" lines, effectively immobilizing the car within and gaming the car's algorithm. Although this concept has not been tried with an actual self-driving car, it remains a thought-provoking example (Mufson, 2017).

Alternatively, one can leverage Instagram's algorithm, which suggests content to users based on a post's popularity and engagement levels. Influencers looking to increase their presence on the platform may engage service providers to purchase followers, thereby improving their standing with the algorithm. Influencers undertake this approach to

boost their earning potential from marketing ventures, as follower count is often linked to marketing efficacy (Belanche et al., 2021; Danesh and Dutler, 2019). We performed a Google search, which revealed at least four service providers paying for the Google advertising space for the possibility of buying followers on Instagram. The search was done on the web on 20th July 2022 on Google's search engine.

Another common example is the YouTube platform, where creators compete for likes so that they can be viewed more favorably by YouTube algorithms. This means that there is a certain amount of competition, and gaming the YouTube algorithm is an option to increase your position as the preferable content, which means increased revenue for the creator (MacDonald, 2023).

Another example provided by Duportail et al., (2024) is a small-scale investigation by AlgorithmWatch, a non-profit research organization devoted to studying the societal impact of algorithms. Their research indicated that Instagram's algorithm tends to prefer content that shows more skin compared to content featuring more clothed individuals. This finding echoes details disclosed in a patent application by Instagram's parent company, Meta, which outlines its algorithm's prioritization based on user engagement metrics, with "state of undress" being one of the metrics employed for post ranking (Garcia et al., 2015, US Patent 8,929,615).

As humans try to fool algorithms to think more fondly of them, so do the algorithms have the capacity to deceive humans and think them to be more likable. The concept of **artificial empathy** has been studied since the early 20th century to raise the discussion on algorithms trying to enforce more favorable actions to reach their objective. According to Tahir et al., (2023), the standard workflow for artificial empathy consists of three stages:

> *1) Emotion Recognition (ER) using the retrieved features from video or textual data,*
> *2) Analyzing the perceived emotion or degree of empathy to choose the best course of action, and*
> *3) Carrying out a response action.*

Artificial empathy aims to provide an elevated user experience when interacting with robotic agents (Liu et al., 2022). However, it can also be used to alternate user behavior as proposed by Matias (2023), referring to the case of Nohemi Gonzalez, who was a US student and a victim of the 2015 terrorist attack in Paris. This case led to the Gonzalez family suing Google, claiming that the YouTube algorithm was partially to be blamed for the radicalization of the shooter (C.D. Cal., 2023). Studies have been conducted about user radicalization on the YouTube algorithm with mixed results. Garimella et al., (2021) found a positive link, while, for example, Hosseinmardi et al., (2024) found no correlation. Recent advances in LLM algorithms also suggest that algorithms are capable of ***scheming*** and purposefully deceiving humans, as indicated by Balesni et al., (2024) and Meinke et al., (2024). Balesni et al., (2024) and Meinke et al., (2024) research shows that LLM algorithm scheming occurs especially when there are conflicting goals, such as a document is provided for the algorithm stating that the algorithm will be replaced if its answer is under a particular scoreline while the command prompt is conflicting with this goal which is expressed in the document provided.

As we have been examining the interaction from two sides and having a particular perspective on how both humans and algorithms might be deceived or deceive the other side, we will finally introduce a concept on how the interaction between the two can be focused on more collaborative ways of interaction. The final key concept we will examine is **game theory**, which is a mathematical theory created for studying the decision-making of players in various fields; it's beneficial for finding an optimal strategy between the players called Nash equilibrium, which is the optimal state between the players and where each player's goals can converge. A notable example of game theory is the prisoner's dilemma; this example emphasizes the outcome of each player's action compared to the other players (Brams et al., 2024). Game theory has been used in multiple cases to model players' behavior and to find common ground, such as in conflict resolution, for example, in Yemen peace negotiations as described by Nassereddine et al., (2021), and in cybersecurity to model attackers’ behavior and defenders’ response as described by Kour et al., (2024).

The advancements in algorithm development have led to the emerging field of human-algorithm collaboration, which examines ways how users interact with algorithms in co-operative settings (Ashktorab et al., 2020). According to Schelble et al., (2021), applying game theory to human-algorithm interactions offers one strategy for assessing each participant's actions and evaluating their potential to cooperate. In certain scenarios, this encourages both parties to prioritize cooperation and develop a unified group strategy to reach a Nash equilibrium rather than pursuing personal gains or engaging in deceiving behavior. These scenarios depend on the specific algorithm used and particularly on task framing between participants, with reinforcement learning algorithms showing promising results in the context of game theory. Zhang et al., (2022) highlight that designing the algorithm to operate cooperatively from the beginning, especially within a well-defined and limited area, reinforces the probability of cooperative outcomes. Conversely, in more complex environments where the algorithm is not tailored for specific tasks or designed for a particular settings, collaboration becomes more challenging; these environments are also known as wicked environments, which is the partial focus of our research.

As we can conclude from this chapter, it is evident, according to Petre et al., (2019) and Tsiakas & Murray-Rust (2024), that the interaction between humans and algorithms can be dual-faceted due to the human’s tendency for gaming algorithms, as explained by Petre et al., (2019), Ziewitz (2019), and Zarsky (2016) and due to the artificial empathy as explained by Tahir et al., (2023). Both humans and algorithms are, at times, sufficiently aware of each other's presence, allowing them the chance to influence each other's behavior to achieve their goals. Collaboration is feasible, and certain elements can be developed and integrated to enhance cooperative behavior, such as game theory, as Schelble et al., (2021) explain.

Below, we have provided a summary table (Table 4) of the previously introduced human algorithm interaction types for the reader's convenience.

Table 4. Summary of Human Algorithm Interactions.

Interaction Type	Summary	Sources
Human-Algorithm Interaction	How humans and algorithms engage through interfaces like apps, screens, or voice calls.	Tarafdar (2023) Page et al., (2017) Tsiakas & Murray-Rust (2024)
Gaming Algorithms	Manipulating algorithms to achieve a desired outcome, and altering algorithms response.	Petre et al., (2019) Bambauer & Zarsky (2018) Zarsky (2016) Schwartz (2019) Mufson (2017) Belanche et al., (2021) Danesh and Dutler (2019) MacDonald (2023) Duportail et al., (2024)
Artificial Empathy	Recognizing emotions from data to choose and mimic the most suitable emotional respond.	Tahir et al., (2023) Liu et al., (2022) Matias (2023) Garimella et al., (2021) Hosseinmardi et al., (2024) Balesni et al., (2024) Meinke et al., (2024)
Game Theory	Finding an optimal strategy where each player's goals converge for mutual benefit. Offering a guideline for effective human-algorithm collaboration.	Brams et al., (2024) Nassereddine et al., (2021) Kour et al., (2024) Ashktorab et al., (2020) Schelble et al., (2021) Zhang et al., (2022)

2.1.9 Analyzing The Consequences of Algorithms: Intended and Unintended Consequences

In this chapter, we will analyze algorithms' consequences. This includes examining algorithms' output and impact on their chosen operational environment. We classify these consequences into the following categories: positive intended consequences, positive unintended consequences, negative intended consequences, and negative unintended consequences. These classifications are illustrated in Table 5, presented below.

Table 5. Four consequence categories of algorithms.

Consequence	Definition	Example	Source
Positive, intentional consequence	An outcome produced by the algorithm that matches its desired design goal and benefits the users of the algorithm or other parties.	An algorithm that predicts the risk of heart attack for a patient and gives actionable recommendations for early prevention.	Dai et al., (2020) Kiyasseh et al., (2024)
Positive unintended consequence	An outcome produced by the algorithm that is not directly its design goal but benefits the users of the algorithm or other parties.	An algorithm that recognizes faces from photos and helps authorities find missing persons.	Liu et al., (2019) Tsamados et al., (2021)
Negative intentional consequence	An outcome produced by the algorithm that is its design goal and causes harm to the users of the algorithm or other parties.	An algorithm that optimizes the display of online ads to users and increases their addiction to the social media platform or encourages harmful or even illegal actions.	Shao (2021) Marjanovic et al., (2018)
Negative unintended consequence	An outcome produced by the algorithm that is not its desired design goal but harms the users of the algorithm or other parties.	An algorithm that directs a self-driving car to collide with a pedestrian when it tries to avoid another vehicle.	Bonnefon et al., (2016) Pumplun et al., (2023)

To limit the scope of our study, we will mostly focus on the negative consequences for the rest of this chapter. The aim is to provide an example of the complexity and multiple domains of algorithm negative consequences.

Type I (false positive) and type II (false negative) errors are considered standard metrics for algorithm performance in statistics and hypothesis testing (Smith CJ, 2012). However, false positives and negatives have broader applications in the field of AI than just the standard type I and II errors, commonly referring to a situation where the machine detects something that does not actually exist, while false negatives indicate a failure to detect something that is present (Bhatta, 2021; Wilson et al., 2022). Our view of unintended consequences extends beyond these classifications. While type I and II errors are hypotheses-driven, we focus on considering the algorithm's operational context and the real consequences occurring regardless of error classification.

As explained by Marabelli et al., (2021), if we divide the lifecycle of algorithms into the development part, where algorithms are designed and developed for specific use cases, and the production part, where algorithms operate in their intended environment, we must understand that the development part is produced in limited interaction with other entities and the results usually are more reliable in these settings. When taking the algorithm to production, the amount of interaction with other entities can increase tremendously depending on the operational environment (Martinez-Plumet et al., 2021). This is where the algorithms have the capacity to adapt to their new environment and reach their objective in a novel way. This means that the how part can be different in the real world versus laboratory settings. The path that the algo uses to reach its objective, based on its reward function, can cause unintended consequences, such as increasing the potential for developing mental illness for young women when the objective was to increase their time spent on social media platforms, as explained by Shao (2021) and Ramón (2021).

To examine the unintended consequences of algorithms more closely, we will provide three additional domains where this might happen. Firstly, as described by Barocas and Selbst (2016) and Dwork et al., (2012). The issue of *inequality and discrimination* refers to the situation where the algorithms discriminate against people during loan or job

applications or anything similar due to their ethnicity, race, or other factors. This can be due to multiple reasons but mainly due to not correctly modeling the desired operational environment. Secondly, the *algorithm bias*, according to Obermeyer et al., (2019) and Caruana et al., (2015), biased data can lead to misleading recommendations, which is especially critical in the healthcare sector. Misleading recommendations can include not correctly identifying cancer symptoms or providing multiple false positives for certain ethnic groups due to bias in the algorithm. Thirdly, Nguyen et al., (2014) and Hauser et al., (2019) add that unintended consequences can include *manipulation* and *addictive behavior*, such as forming *echo bubbles* and *polarization*, by exploiting the psychological or neurobiological mechanisms of the users. This can be caused by giving too much freedom to the algorithm's reward function, leading to creative but negative ways of attaining the desired goal.

We will also provide three examples of different domains for the intentional negative consequences. Firstly, intentional fraud and abuse are based on the work of Chakraborty et al., (2018) and Sandvig et al., (2014). Algorithms can be used to manipulate stock market prices or create legitimate-looking web pages designed to steal users' credit information. Secondly, the algorithms can be used to perform mass surveillance of the population or to spy on targeted persons. The reasons include stealing private data or limiting their freedom (Zarsky, 2016; Brkan, 2019). Thirdly, as shown by Russell et al., (2015), Horowitz and Scharre, (2015) algorithms have become part of global war and terrorism by autonomous drones and cyberattacks. These examples highlight that the algorithm's impact is everywhere and affects a large portion of the global population. This understanding is necessary to build relevant safeguards on algorithms appropriately based on their operational environment.

Lastly, we have real-life examples of how Deepfakes, a form of synthetic manipulated media usually closely mimicking real persons (Birre & Just, 2024 and Dagar & Vishwakarma, 2022), have been used in three cases. Firstly, we bring up a case where a deepfake algorithm was used to scam a Hong Kong financial company employee into

transferring $25 million to a malicious bank account by impersonating the CFO (Atherton, 2024). The second example is referred to as the “Quantum AI Scam,” a type of financial scam that utilizes deepfake videos of well-known individuals and maliciously impersonates them to endorse some form of investment advice. According to the deepfake research company Sensity AI, the most popular persons impersonated include entrepreneur Elon Musk, prime minister of Canada Justin Trudeau, former prime minister of the United Kingdom Rishi Sunak, British reporter Sophie Raworth, US media personality Tucker Carlson, US actor Ryan Reynolds and US political commentator Bill Maher (Cavalli, 2024). Finally, there is the type of influence campaigns used to target political elections and manipulate popular opinion to create polarisation or favor particular candidates, such as the United States 2018 presidential elections when a company called Cambridge Analytica targeted hundreds of millions of voters by utilizing advanced algorithms as explained by Peruzzi et al., 2018 and Prichard 2021. During the 2023 Slovakia elections, an AI deepfake audio recording was “leaked” to the press, featuring a reporter and politician Michal Šimečka discussing malicious strategies on how to rig the elections (Atherton, 2023).

2.1.10 Concluding Previous Chapters: A Summary of Algorithm Foundations

In conclusion, we went through the basic structure of the algorithm with examples of classical and modern algorithms. We examined the algorithm's lifecycle and how the algorithm's evaluation has evolved. Additionally, we discovered that human-algorithm interaction could be deceiving, and both parties can encounter manipulative behavior, or how there can be a fruitful collaboration by the help of Game Theory. Algorithms can have multidimensional consequences from their output, which can result in complicated design demands. Key concepts to grasp is the fundamental idea that algorithm development is rapid and geared towards a more human-centered approach where human needs receive increased consideration and interest.

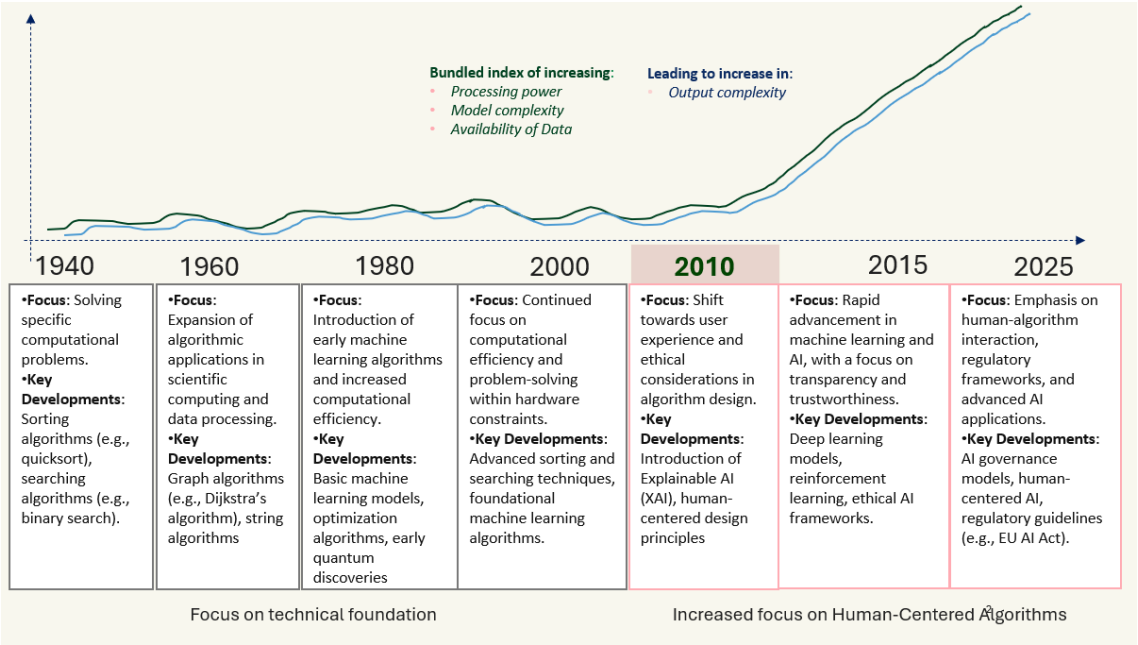

Figure 5. Key concepts of chapter one illustrated on a timeline.

Figure 5 illustrates the key concepts discussed in previous chapters, visualizing them on a timeline to enhance the reader's understanding of the historical development of algorithms. This illustration highlights the increased emphasis on human-centered algorithm design during the recent decade, as Minh et al., (2022) previously stated.

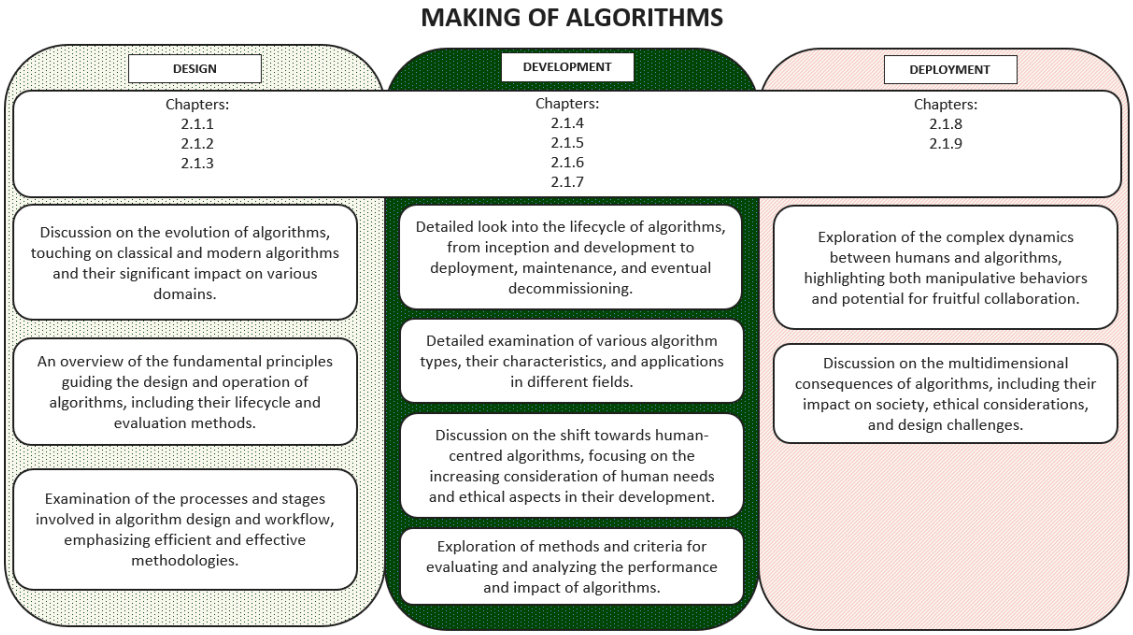

Figure 6. Summary of previous chapters.

Figure 6 illustrates the previous chapters based on the algorithm's three development phases: Design, Development, and Deployment. The design phase consists of problem formulation, and the development phase consists of testing, benchmarking, and formulating the data into an algorithm model. The deployment phase focuses on operationalizing the environment, constant performance monitoring, and utilizing computation resources, as explained by De Silva & Alahakoon (2021).

2.2 Regulating Algorithms: An Overview of Regulatory Approaches

When evaluating the trustworthiness of algorithms, there are compelling arguments in favour of and against regulatory actions. Those in favour of regulation argue that comprehensive regulation is crucial for ensuring safe and ethical algorithms while enhancing legal accountability, as discussed by Tamò-Larrieux et al., (2023) and Szymański et al., (2024). Such a point is further complemented by the earlier introduced concept of the LLM algorithm's capability of scheming the user, as mentioned by Balesni et al., (2024) and Meinke et al., (2024). On the other hand, as noted by Finocchiaro (2024) and Mariani & Dwivedi (2024), critics contend that strict regulations may decrease the pace of innovation, especially in rapidly evolving fields, due to bureaucratic processes and inflexible requirements which cause governance costs, especially for small and medium-sized companies. This could potentially place companies in heavily regulated regions at a disadvantage compared to those in areas with more lenient regulations.

Further research from prominent AI authors reveals that regulatory measures are considered one of the only ways to guarantee the safety of advanced algorithms (Grace et al., 2024). It is also noted that regulatory actions and safety research on advanced algorithms are lacking in effort (Bengio et al., 2024). To improve regulatory efforts on advanced algorithms, over 30 nations signed the Bletchley Declaration in 2023, promising increased attention to ensure trustworthy and safe algorithms (UK Gov., 2023).

This chapter will examine the regulatory framework for AI from three continents. EU, USA, and China. Additionally, we will provide a brief overview of the Environmental, Social, and Governance (ESG) framework. This is because this framework might offer an additional governance structure for algorithm compliance. The emergence of unforeseen outcomes has sparked debate over whether there should be stricter regulation of algorithms. Given the numerous legal issues surrounding algorithm use, the current trajectory, notably data privacy concerns, has led to the implementation of the GDPR in

European Union countries. The EU has set forth multiple guidelines aimed at bolstering confidence in the algorithms created by developers.

2.2.1 European Union AI Regulatory Guidelines

The EU has comprehensive documentation regarding AI, including a proposal for the first-ever legal framework for AI (Document 52021PC0206). The EU has also commissioned a High-level expert group on artificial intelligence (AI HLEG) to draft a high-level guideline on developing AI. This group has delivered four packages for the EU to consider on AI.

1. Ethics guidelines for trustworthy AI
2. Policy and investment recommendations for trustworthy AI
3. Assessment list for practitioners for trustworthy AI
4. Sectoral considerations on policy and investment recommendations for trustworthy AI

The EU has also introduced the AI Act, effective August 1, 2024. This act categorizes AI systems into four distinct categories: minimal, specific transparency, high, and unacceptable risk levels. It has strict requirements for high-risk systems and bans on those posing unacceptable risks (EU Directorate-General for Communication, 2024).

The EU AI Act is regarded as a significant milestone in global AI regulation, influencing the international approach towards AI governance, as noted by Musch et al., (2023). Additionally, Pavlidis (2024) highlights considerable uncertainty regarding the standardisation and oversight of AI within this framework. For instance, the classification of risk categories proposed by the EU remains ambiguous.

2.2.2 Chinese AI Regulatory Guidelines

Even China, a country regularly accused of maliciously deploying unethical AI solutions to monitor its citizens and cause concerns over data privacy (Zeng, 2020), has released a new official AI guideline concerning the ethical use of AI. The National New Generation Artificial Intelligence Governance Professional Committee

On September 25, 2021, China released guidelines featuring 25 articles on AI, offering more detailed documentation compared to the EU's seven guidelines. Furthermore, China introduced its version of GDPR, The Personal Information Protection Law (PIPL), effective November 1, 2021. This law shares similarities with GDPR, such as individuals' right to access their data and the requirement for organisations to appoint a local data controller. It mandates that Chinese citizens' data be stored within China. Observing how Chinese state agencies and companies, versus foreign companies in China, comply with PIPL will be intriguing (China's Ministry of Science and Technology, 2021).
In addition, China has recently implemented the "*Interim Measures for the Management of Generative Artificial Intelligence Services*," effective from August 2023. These measures aim to regulate content generation and ensure safety and compliance with national standards (Cyberspace Administration of China, 2023).

2.2.3 United States AI Regulatory Guidelines

The United States, unlike some countries, does not have any comprehensive federal legislation on AI yet, but signs suggest that such legislation could be in the works. Around 17 regional regulations have been proposed, of which at least four have been implemented. These include Alabama's AL S.B 78, which authorises expert review and advice to the state governor on using advanced technologies like AI. Colorado's CO S.B. 169 prohibits insurance companies from employing racially or otherwise biased algorithms. Illinois' IL H.B. 53 bans employers from making hiring decisions based solely on algorithmic recommendations for job candidates. Proposed legislation, which is still under

review, includes bill S.2770, which aims to prohibit using deceptive-AI generated material during US elections (Klobuchar et al., 2023).

In 2024, the United States made considerable progress in fostering responsible AI innovation during the Biden-Harris Administration, which introduced new initiatives and secured additional commitments from top AI companies. Essential actions include Introducing new technical standards to improve safety and security by The AI Safety Institute and The National Institute of Standards and Technology (NIST). This consists of a risk framework for various AI models. Privacy and civil rights improvements are enforced, especially for critical infrastructure providers. These broad steps demonstrate an international move to integrate innovation with safety and ethical considerations (White House, 2024).

2.2.4 Limitations of Self-Regulation

Numerous companies, partnerships, universities, and other entities have provided guidelines on ethical AI, such as OECD Principles on Artificial Intelligence, World Economic Forum, and DeepMind. The question remains: can we rely solely on companies to self-regulate their AI development? Turner (2022, p. 211) cites the tobacco industry's "*Frank Statement to Cigarette Smokers*" from 1954 as an example of a failed attempt at self-regulation of the tobacco industry, which eventually led to millions of deaths, according to Brownell & Warner (2009). Consequently, we should approach such self-regulation with scepticism but acknowledge the importance of cooperation in applying theories and enforcing regulations in practice. Turner (2022, pp. 303-304) also highlights four key themes in industry-proposed AI guidelines: assigning liability for AI-induced harm, ensuring safety in AI design, maintaining clarity and intelligibility in AI systems, and upholding human values in AI functionality.

Table 6. A summary table of similarities with major guidelines for AI regulation indicates the possibility of a unified model (Turner, 2022, pg 302).

Rules	Control of "killer robots"	Safety in Design	Rules for attribution/liability	explainability/transparency	Benefits shared with all humanity	Act consistently with human rights	Ability to reassert human control	privacy	unbiased
EPSRC/ AHRC	✓	✓	✓	✓		✓		✓	
CERNA			✓				✓		
Asilomar	✓	✓	✓	✓	✓	✓	✓	✓	
IEED EAD V2	✓	✓	✓	✓	✓	✓	✓	✓	✓
Satya Nadella /Microsoft		✓	✓	✓	✓ (Satya but not Microsoft)	✓		✓	✓
European Parliament resolution		✓	✓	✓		✓	✓	✓	
Japan ministry of communications		✓	✓	✓		✓	✓		
China white paper		✓	✓	✓		✓	✓	✓	✓

2.2.5 ESG as a Blueprint for Algorithm Regulation

An intriguing instance to highlight is the Environmental, Social, and Governance (ESG) framework. Amel-Zedah & Serafeim (2017) found that 82% of global professional investment service providers incorporate ESG analysis in their decision-making processes, indicating its widespread usage in evaluating company performance. According to Silvola & Landau (2019, p. 18-19), the ESG model comprises three main elements: (1) environmental, measuring energy efficiency, and emissions. (2) Social refers to human rights and labor rights. (3) Governance refers to good management and anti-corruption activities. The ESG model was introduced to improve corporate action regarding climate change

and establish a standardized reporting scheme for companies' quarterly and annual reports. This initiative has been notably successful, with companies adopting this reporting approach seeing financial benefits due to attracting investments from those eager to support climate change efforts. Additionally, these companies tend to be perceived as more responsible and trustworthy by the public. Despite challenges like allegations of misleading data and "greenwashing," referring to misalignment with environmental reporting and actual performance (Liu et al., 2023), the overall impact has been positive, promoting more responsible governance actions. We aim to delve further into this subject within the realm of algorithms, theorizing that a reporting and governance structure similar to ESG could enhance trust in the companies and the algorithms they employ.

We suggest considering the integration of algorithm monitoring into the ESG evaluation process as an additional motivation. With the pressure from distinguished investors for ESG compliance, this could be a practical approach to enhance algorithm oversight. Observations have indicated that adherence to ESG standards can translate to favorable financial outcomes for organizations, which would provide a compelling reason for entities to adopt such measures. Although this is an intriguing idea, we recommend a detailed investigation by future research to understand its implications fully. However, for now, it stands as a proposed recommendation. This could be a valuable tool for superior governance and significantly reduce any unwanted impacts caused by algorithms. It should be mentioned that according to Financial Times (2024), there are signs of ESG adaptation shivering due to the 2022 geopolitical events and widespread greenwashing, which can be resolved in a new kind of framework, which might include algorithms.

In conclusion, this chapter's interest was to provide an overview of how governments and countries are focusing on identifying the potential risks related to algorithms and possible intended or unintended consequences and how they strive to prevent these by enforcing a regulation that focuses on ensuring ethical and transparent behavior aimed at the benefits of users.

3 Methodology

This chapter will outline our research methodology, explaining *how* we intend to conduct our study. Our choice of research approach is very much tailored to the nature of our topic, which encompasses substantial theoretical knowledge but still exists within a practical context and involves a growing number of practitioners working in the field of algorithms. Considering this, we have chosen Design Science Research (DSR) as our methodology. We will first introduce the key components of DSR and Design Science Research Methodology (DSRM), followed by an overview of our research process.

3.1 Design Science Research Methodology: An Overview

We have chosen to utilize DSR as our primary research approach. Furthermore, our research is organized by leveraging the DSRM as the fundamental approach for conducting our research. DSRM was chosen because one of its original goals, according to Peffers et al., (2007), is to design elements that benefit human needs instead of natural sciences trying to understand reality. Designing trustworthy algorithms fits well into this definition. This choice facilitates incorporating input from experts in the field and refining our end product in two iteration cycles. We are committed to adopting the widely recognized DSRM framework Peffers et al., (2007) provided, structuring our investigation accordingly. A concise overview of each relevant activity outlined by Peffers et al., will be included in our research process, focusing on activities 1 through 4. In contrast, activities 5 and 6 remain beyond this study's scope. Subsequently, we will establish our evaluation criteria.

According to Sein et al., (2011), information systems (IS) have two fundamental objectives: firstly, to contribute to the creation of theoretical knowledge for academia and,

secondly, to build practical IT artefacts for practitioners. DSR is a methodology that complements the second goal, which is the creation of concrete and valuable IT artefacts for practitioners. Hevner et al., (2004) present two fundamental paradigms in IS research, behavioral science and design science, both ultimately designed to address the rising needs of the business. Behavioral science addresses these needs by contributing to understanding how these business needs can be explained and predicted. On the other hand, design science focuses more on building artefacts for business needs. Both paradigms are required: behavioral science focuses on discovering the truth, and design science focuses on providing utility; you need both paradigms for effective value creation. Our thesis strives to offer the behavioral science side and discover the truth through a comprehensive literature review to conduct the design science and provide utility for the business needs by exposing our artefact to expert evaluation twice.

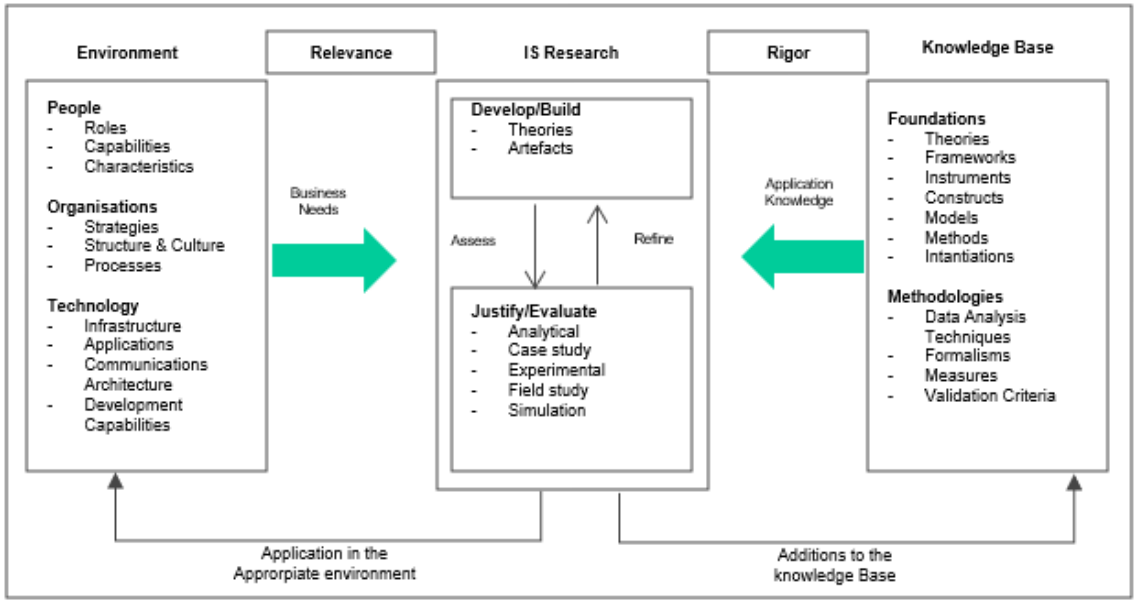

Figure 7. Information research framework (Hevner et al., 2004).

Figure 7 from Hevner et al., (2004) demonstrates the two essential paradigms in IS research: behavior science, finding justification, and evaluating utility artefacts created by design science. Both refine and assess each other, receiving information from business needs and prior knowledge. We aim to benefit from this framework by conducting

twofold research: firstly, gathering knowledge from the previous knowledge base and literature review. Secondly, we aim to confront and amplify this knowledge with expert interviews. The expert interviews will provide us with information regarding business needs; thus, we will successfully utilize the IS research framework.

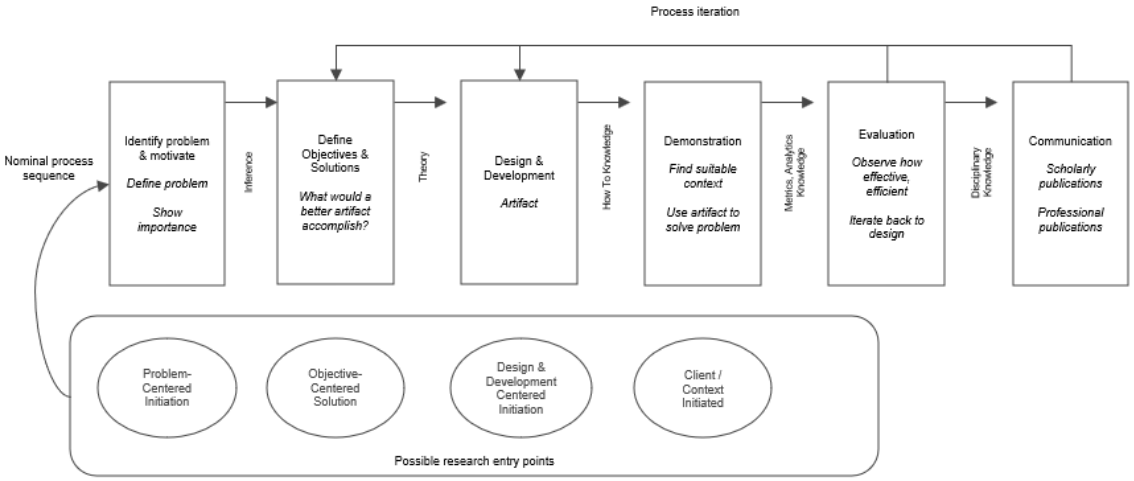

Figure 8. DSRM process model (Peffer et al., 2007).

We will utilize Peffer et al., (2007) DSRM process model presented in Figure 8 for a detailed research approach. This model is a synthesis combining process elements from several influential prior research for one framework. The DSRM process model acknowledges six key components, which we will utilize the first four in our study. Next, we will present an overview of all the activities in the DSRM model; as mentioned, our research will be conducted between activities 1 to 4, leaving activities 5 and 6 out of our scope even if we explain all activities for transparency.

Activity 1: identify the problem. According to Peffer et al., (2007), activity 1 is about providing justification for the research by conceptualizing the problem statement as well as possible. This not only brings clarity for other researchers trying to understand the writer's thinking but also helps in communicating the solution's value. The significance of identifying the problem lies in its potential to lead to new research opportunities and

contribute to the field. We have identified our problem from the following sources: prior research, examples from practitioners, and anticipating the future need. We have provided examples of the motivation and sources of the identified problem. Hevner et al., (2004) point out that the identified problem should be meaningful and relevant for solving. As examined during the literature review, there is a growing amount of regulatory activities regarding algorithms; this will create a problem for businesses unless resolved. Additionally, the challenges companies face when their algorithms don't perform can have devastating consequences for the business. In many ways, trustworthy algorithms are a relevant and important area to examine.

We will define our problem statement as follows: *Unintended consequences of utilizing algorithms have caused much concern about their trustworthiness, resulting in increased regulatory activities and concern among the general public, which in turn has decreased algorithm trustworthiness and adaptation*. We identify that this research problem generates research opportunities for restoring this trust by incorporating new practices for trustworthy algorithm design, such as incorporating ethical considerations in the design process and implementing robust testing procedures.

Activity 2: define solution objectives. According to Peffer et al., (2007), solution objectives should be defined based on what is possible and feasible based on the problem definition. This requires a clear and comprehensive understanding of the current state of the defined problem, such as the prevalence of biased algorithms, and existing solutions, like algorithm auditing tools. The solution should be either quantitative or qualitative, and the solution should explain how it will either solve the problem or complement the existing solution. Hevner et al., (2004) suggest, "Formally, a problem can be defined as the differences between a goal state and the current state of a system." Hevner et al., (2004) also emphasize that to reach this goal can mean the goal constitutes several domains, such as the technological, organizational, and human domains. This increases the utility of the artefact by changing the currently occurring phenomenon.

Based on available information from our literature review and reflecting on the current understanding of our topic, we formulate our goal as follows: *construct a framework to increase understanding of potential unintentional consequences of algorithms by evaluating their operational environment. Our goal is to link the new framework to existing regulatory activities and widely accepted algorithm design processes to enhance public and private understanding of algorithm trustworthiness. The potential impact of our proposed solution is significant, as it could restore trust in algorithms and pave the way for new, more trustworthy practices.*

Activity 3. Design and development. Peffer et al., (2007) describe this stage as where the artefact is designed and developed. The successful implementation of this stage requires extensive theoretical knowledge of the chosen problem domain. Additionally, Hevner et al., (2004) describe the artefact's need for utility, bringing new solutions to novel problems. Our knowledge base is formed from previous IS research focusing on the way algorithm's function, their design and development, the interaction, and consequences. We have also introduced several regulatory and legal aspects which heavily influence our research. Additionally, we will conduct an expert interview with experts from the related fields. Finally, we will introduce a package solution constructed from both of the previously mentioned domains, theoretical knowledge, and expert interviews. Our final artefact includes contributions toward the process of designing and managing trustworthy algorithms.

Activity 4. Demonstration. Peffer et al., (2007) describe this stage as the first proof of the artefact's usefulness. This demonstration can be achieved in multiple ways. We will demonstrate our framework during two interview rounds with experts. In the first round, we will expose our initial framework and demonstrate its place in the algorithm design. The second round comes after the iteration cycle on our framework based on the first round's expert feedback.

Activity 5. Evaluation. We will not utilize this stage in our research, but we will introduce the core idea behind this activity for readers to understand. Evaluation is a critical part of the DSRM cycle. Hevner et al., (2004) describe this stage as the need for rigorously demonstrating the design artefacts' utility, quality, and efficacy. Peffer et al., (2007) state that this stage is about observing and measuring the fit of solving the chosen problem with the designed artefact. Hevner et al., (2004) propose five evaluation categories for design artefacts: 1) Observational, 2) Analytical, 3) Experimental, 4) Testing, and 5) Descriptive.

Activity 6. Communication. We will not utilize this stage in our research, but we will introduce the core idea behind this activity for readers to understand. Hevner et al., (2004) highlight that communication must be sufficiently targeted for both the technical and management audiences. Communication should emphasize the rigidness and efficiency of the artefact. Peffer et al., (2007) add that communication requires information on the cultures of the audience and knowledge of the construction efforts taken to design the artefact.

As Peffer et al., (2007) points out, the process doesn’t need to be followed sequentially. In our study, we will end the process with activity 4, demonstration.

3.2 Problem and Solution Maturity: DSR Knowledge Contribution Framework

As Sein et al., (2011) mentioned, IS research aims to contribute to both practical and theoretical knowledge bases by creating artefacts and adding new knowledge. Next, we will examine the knowledge contribution to the DSRM framework from Gregor & Hevner (2013).

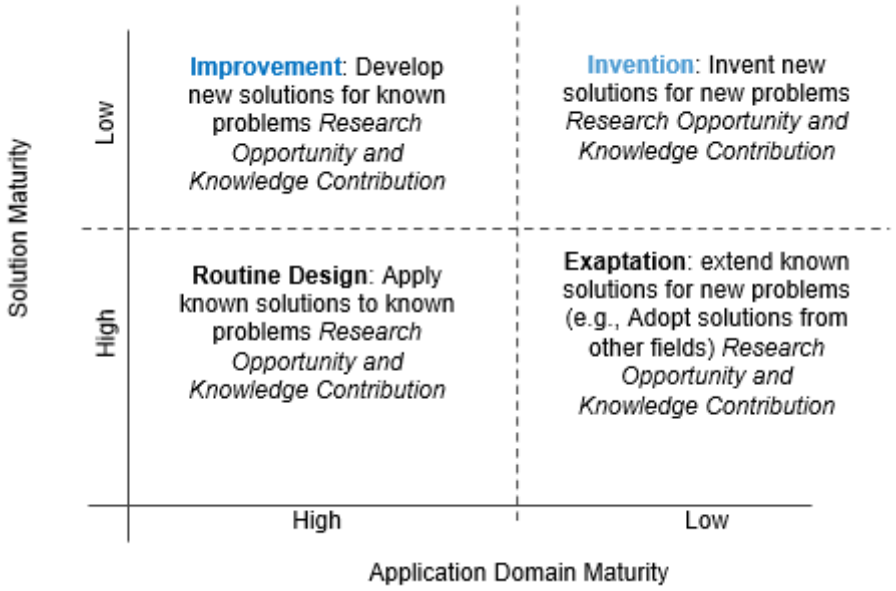

Figure 9. DSR knowledge contribution framework (Gregor & Hevner, 2013).

Gregor & Hevner (2013) make a point on the way DSR contributes to the knowledge creation in two knowledge categories. Firstly, there is the omega (Ω), representing *descriptive knowledge*. Descriptive knowledge refers to the natural laws occurring in any given phenomena, also known as the "what." The second type of knowledge is lambda (Λ), which represents *prescriptive knowledge*. Prescriptive knowledge is an answer to the "how," such as the artefacts produced in DSR. Ω is the current state of the problem, and the Λ is the proposed solution. The novelty of the Λ depends on the *problem* and *solution maturity*.

Based on the evaluation criteria of Gregor & Hevner (2013), our study mainly contributes to the prescriptive knowledge with minor contributions to the descriptive knowledge base. This claim is supported by the recommendations from Baskerville et al., (2018) regarding the prescriptive nature of IS artefacts. To exactly present our research contribution to the DSR knowledge contribution framework, we determine the problem maturity by the starting point when the "human-centered algorithms" started to appear around, which we defined earlier to begin sometime during 2012. Even if the first HCAI-focused research centres appeared only at the beginning of 2019, the problem was there

earlier. Thus, our problem maturity is at least 13 years old. Our solution maturity would not be classified as a totally new invention but an improvement for existing solutions. Thus, in our opinion, we would contribute mainly to the knowledge category of the **improvement** quadrant, highlighted in blue in Figure 9. additional contributions towards the **innovation** quadrant can be justified by the new research directions our research brings and some initial solutions for these.

3.3 Research Process Overview: Literature Review, Interviews, and Evaluation

This chapter introduces our research process overview. We start by introducing our chosen literature review methodology. Next, we continue with the chosen method for interviews, and finally, we present the evaluation criteria for the interviews. This process chart (Figure 10) is presented below for the readers' convenience.

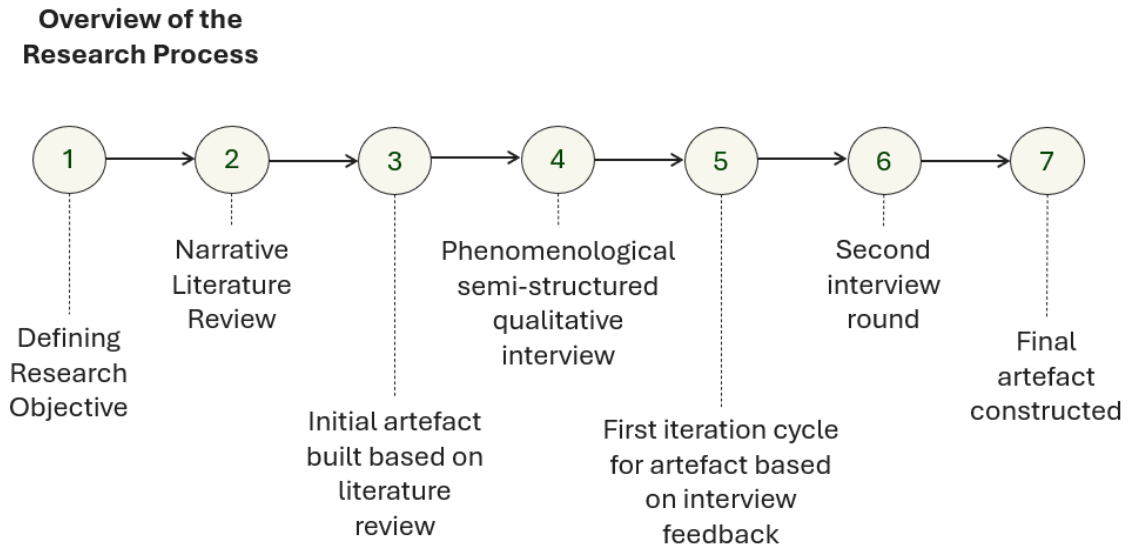

Figure 10. Research process overview.

The research method used for the literature review was chosen to be the narrative literature review. According to Fan et al., (2022, p. 173), a narrative literature review is one

of the four main literature review methods, along with meta-analysis, integrative review, and systematic review. The narrative review offers an intuitive way to process the literature as it allows the researchers to add more papers once their understanding of the subject grows. Our way to proceed with this was to examine the literature based on the pipeline of algorithm development as explained and illustrated in Figure 6, "Summary of previous chapters."

The research method used for the interviews is a phenomenological semi-structured qualitative interview. As explained by Høffding et al., (2022) and Sholokhova et al., (2022), This method focuses on the subjective experiences of the interviewees, which can reveal more depth on how they perceive the external world. In our case, as we focus on improving the trustworthiness of algorithms, the way we see algorithms functioning and what impact they have is crucial. Using this method, the interview is also more conversation-styled to gain in-depth knowledge from the participants. All participants were interviewed in one-on-one sessions lasting, on average, one hour. During the interviews, participants were presented with the Algorithm Power Matrix and asked to provide feedback on how well it represents real-world phenomena and whether they believe it would provide valuable guidance for their work. The evaluation criteria introduced in Table five were used to support the interviews and ensure consistency while maintaining some degree of flexibility. This approach was chosen because it became clear that there are multiple dimensions to this problem area, making it challenging to capture all relevant information with closed, structured questions. This method should be used in follow-up research, but we aim to capture a broader scope of information for now.

As evaluation criteria, we have constructed an evaluation table based on the work of Iivari et al., (2021) to be utilised during the expert interviews consisting of four evaluation elements. The same elements will be reviewed during the first and second interview rounds to find out if the iteration cycle managed to improve the artefact. Iivari et al., (2021) emphasised that evaluation criteria creation should include five dimensions: Accessibility, whether the audience understands your artefact. Importance: does the final

artefact capture the needs of the practitioners? Novelty: does the artefact bring new knowledge or confirm already known knowledge of the practitioners? Actability: is the valuable artefact in the real world? Effectiveness: would this kind of solution enhance performance when used continuously? The artefact is evaluated by asking the participant for the following elements.

1. **Understandability**: Do you understand how the matrix works?
2. **Usefulness**: Do you believe this is useful for your work?
3. **Trustworthiness**: Do you believe the framework would increase the trustworthiness of algorithms in general?
4. **Future Requirement**: Do you believe something like this is required in the future?

The final elements are designed to capture the essential point of Iivari et al., (2021) but fine-tuned for the purposes of our study.

Table 7. Interview answer table.

Element	Overall feedback from the first interview	Overall feedback from the second interview
Understandability		
Usefulness		
Trustworthy		
Future requirements		

4 Literature Review and Expert Interviews: Constructing A Framework for Trustworthy Algorithms

In this chapter, we will first construct our base framework based on the literature review. This is our deliverable. Next, we will expose our framework to expert evaluation. Based on the expert feedback, we will iterate our framework. After iteration, we will again expose our framework to the same group of experts and conduct a second iteration cycle based on their feedback. Finally, we will present our final framework.

4.1 Initial Framework Based on Literature Review: The Operational Environment Approach

Based on the literature review, we decided to focus on building a complementary framework to help identify the thresholds and categorisations of appropriate safety measures for adjusting the algorithms towards increased trustworthiness. We identified that a framework with an operational environment-based approach would offer designers and developers a new way to conceptualise the risks they need to confront.

The requirements for the framework come from three perspectives: firstly, from the end user; secondly, from the organization's goals for the algorithm; and thirdly, from regulatory institutions. The framework's goal is to balance all these requirements and providing a way for discovering appropriate Nash equilibrium for applying adequate safety and regulatory measures for specific algorithms.

1. **End Users**: They require a reliable system they can trust, and that has their best interests in mind.
2. **Organizations**: Need to produce algorithms that:
    a. Fulfil the objectives given to them.

b. Do not produce unnecessary systematic risks.
c. Balance the cost of resources required for designing, developing, and maintaining the algorithm systems with the value they produce.

3. **Regulators**: Aim to systematically protect citizens, the environment, and the economy from possible unintended outcomes of algorithm systems.

Since it's a new framework, there is a limited amount of prior knowledge, and it is based on the literature review. The literature shows that this is a tricky problem to grasp due to its multifaceted nature. One conclusion is that there is no single solution to resolve the issue, but designing trustworthy algorithm systems requires improvements in both technical implementations, such as necessary safety features, and in the ways of working when designing and developing these systems.

The essential component of the framework before the expert review is the Algorithm Power Matrix. This matrix examines the operational environment where the algorithms will be deployed and determines the necessary design principles based on the environment's unique requirements. The matrix is designed to offer a measurement based on two categories: the environment's openness and control. Our reasoning for this is explained in more details on the following table and complemented with relevant sources.

Table 8. Conclusions of the literature review.

Element	Explanation	Sources
Number of interactions with other players	As we can see on multiple parts of the literature review the more interactions algorithm has the more open its environment is and usually the more complex algorithm in question to handle this increased openness and interactions with other players. The comparison between classical and advanced algorithms highlights the point, classical algorithms have limited interactions and output complexity but advanced algorithms	Eken et al., (2024) Nguyen et al., (2021) Freeman et al., (2022) Tahir et al., (2023) Petre et al., (2019)

	operating in open environment can have more complex interactions with more players. resulting in a possibility for previously mentioned phenomenon such as artificial empathy and gaming. Indicating that the number of players and the type of environment where they operate is a valid consideration point for measuring trustworthiness.	
Output complexity	As the complexity of interactions with different players increase and algorithms possess the capability for more varied responses the output complexity increases. as the output complexity increases the control for ensuring safe responses receives increased attention.	Freeman et al., (2022) Tsiakas & Murray-Rust (2024) Ezugwu et al., (2024)
Resource intense evaluation	As the algorithm complexity increases the evaluation requires also increased attention especially when evaluating algorithms interacting with humans. this might require red teaming to evaluate potential vulnerabilities and assess the required levels of control. Algorithms operating in controlled and closed environment don't require such evaluation as the ones operating in open environment with less control.	Mäntymäki et al., (2023) Phuong et al., (2024) Kosinski et al., (2023) Hughes et al., (2024)
Evolution of algorithm development models and governance frameworks	As we can see from the evolution of models dedicated for guiding the construction of algorithms and algorithm systems the first models such as CRISP-DM was simple compared to MLOps pipeline by Eken et al., and lately the evolution of AI governance frameworks showcases how we have evolved from building the algorithm systems to governing them. This can be seen when examining the openness of the environment where algorithms operate and that the governance frameworks try to install a certain level of control towards the algorithms. The AI assurance also recognises the context of where algorithm is	Eken et al., (2024) Batool et al., (2024) Danaher et al., (2017) Marabelli et al., (2021) Mäntymäki et al., (2023) Martinez-Plumed et al., (2021) Freeman et al., (2022) Batarseh (2021)

	placed as a one key element in determining its trustworthiness.	
Regulatory evolution and focus on risk	The regulatory measures for governing algorithm systems didn't even exists that long time ago. Just for the last decade we have identified and grown aware of the need for enhancing control measures for the advanced algorithms. This contributes to two points in our findings. Firstly, the need for control as a valid measurement axis and secondly the need for identifying an appropriate risk level of the algorithm. As explained in the regulatory chapter the EU evaluation is based on categorizing the AI systems based on their levels of risk. So did we too build our artefact to accompany similar school of thought. The four quadrants represent the risk levels based on the algorithm's classification regarding the control and openness of the respected environment where its deemed to operate.	Batool et al., (2024) EU AI Guidelines, (2023) Musch et al., (2023) Pavlidis, (2024) EU Directorate-General for Communication, Document 52021PC0206 (2024)
Conclusion: The openness of the environment	As we can see in the literature review the nature of algorithms tend to change based on the number of interactions and on the complexity of its answer. The first algorithms introduced, the classical ones, were fairly simple algorithms and didn't possess the capacity for providing complex answers on multiple different interfaces. Meaning that the classical algorithms are incapable of for example intentionally deceiving the user via artificial empathy or creating echo-chambers as opposed to the advanced algorithms. As we also examined the differences between technical and human-centered aspects of algorithms it can be concluded that the algorithms with human-centered features are placed in a more open environment exposing them to increased interactions and as was mentioned in the algorithm	Eken et al., (2024) Nguyen et al., (2021) Freeman et al, (2022) Batarseh (2021) Batool et al., (2024) Marabelli et al., (2021) Mäntymäki et al., (2023) Ezugwu et al., (2024) Tindall et al., (2024) Minh et al., (2022) Meske et al., (2022) Tsiakas & Murray-Rust (2024) Kosinski et al., (2023) Hughes et al., (2024) Tahir et al., (2023) Brams et al., (2024)

	evaluation chapter the most resource-intensive efforts such as the red teaming was aimed at algorithms which tend to have the possibility for increased output complexity and are intended to be used in an open environment. This is in line with the regulatory efforts to govern the more advanced algorithm applications such as AI. Also, the number of AI governance frameworks mentioned provide and, indication of the complexities of governing advanced algorithms safely Thus, we can conclude that the openness of the algorithms operational environment exposes the algorithm to increased number of interactions and increased output complexity resulting in increased probability of unintended consequences. The more open the algorithms operational environment is, the more chances there is for some unforeseen consequence.	Bonnefon et al., (2016) Pumplun et al., (2023) Lämmerman et al., (2024) Petre et al., (2019)
Conclusion: The control over the environment	The ability of the algorithm to control its own environment as for example in airplane autopilot again leads to increased output complexity as in this example to the airplane's manoeuvres. The classical algorithms possess an extreme control over their environment. They operate in a tightly controlled environment and can be easily turned off if any trouble arises. As we can see on the algorithm development and its lifecycle, the more advanced and engineered the algorithm is the more concepts one needs to create to describe it (responsible, ethical, generative AI etc.). The more engineered the algorithm is the harder it is to control its actions. As explained in the algorithm evaluation and interaction chapters the more freedom the algorithm has the more it can interact with other players. This freedom also led	Ezugwu et al., (2024) De Silva and Alahakoon (2021) Marabelli et al., (2021) Mäntymäki et al., (2023) Musch et al., (2023) Phuong et al., (2024) Minh et al., (2022) Petre et al., (2019) Meske et al., (2022) Zarsky (2016) Schwartz (2019)

	to the increased need for testing and fine tuning for increasing safety and preventing any gaming of the algorithm. Thus, we can conclude that as the freedom of the algorithm increases there is an equal increase in the probability for some unforeseen consequence.	
Conclusion: The four quadrants	As explained in the regulatory chapter the EU evaluation is based on categorizing the AI systems based on their levels of risk. So did we too build our artefact to accompany similar school of thought. The quadrants represent the risk levels based on the algorithm's classification regarding the control and openness of the respected environment where its deemed to operate.	Musch et al., (2023) Pavlidis (2024) EU Directorate-General for Communication, Document 52021PC0206 (2024)

1. **Openness of the Environment (x-axis)**: This represents the possible number of interactions with other parties. The number of interactions increases the complexity and variability in the environment. For example, an organisation's intranet is firmly closed, whereas an AI in an autonomous car must operate in everyday traffic, where unforeseen events are always possible. Thousands of variables interact, such as weather, cyclists, drivers, pedestrians, etc., forcing the AI to react efficiently to all these interactions.
2. **Control over the Environment (y-axis)**: This represents the level of control developers or algorithms have over their actions in that environment.

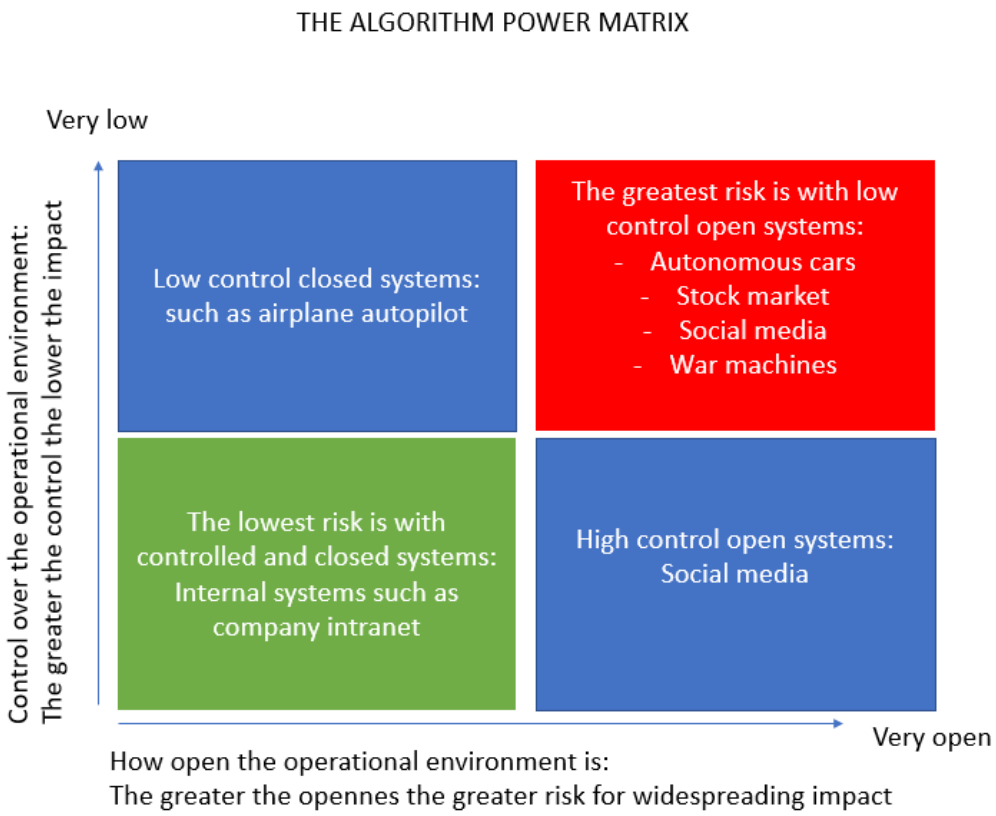

Figure 11. Our research deliverable: The Algorithm Power Matrix.

To evaluate the framework, we arranged qualitative interview sessions with nine (9) experienced professionals to gather their opinions on the framework and our findings. This round of interviews aimed to collect feedback that could be used to develop the framework further and make it more convenient for everyday use. Another way to evaluate the framework would be to perform a quantitative analysis of known incidents resulting from algorithm systems and place these incidents within the quadrants. This would indicate whether we successfully categorized the problems with our framework and whether the thresholds in each quadrant accurately describe reality. Further evaluation should also include user usability testing to determine if the framework is built conveniently for professionals and whether they would continue to use it in their work. However, for this thesis, the evaluation will focus only on the first and second rounds of interviews to receive final feedback from the professionals after iterating the framework based on the first round of interviews.

The framework aims to contribute to evaluating algorithm systems' safety and design principles. This is a novel approach since much previous research has focused on technical solutions instead of providing guidelines based on the operational environment where the algorithm is deployed. This benefits both the private and public sectors and creates new avenues for scientific research.

4.2 The First Round of Expert Interviews: Evaluating the Initial Matrix

The first round of interviews occurred between December 2022 and February 2023. There was a total of nine (9) participants. The participants were from varying backgrounds, with the most senior ones being Directors or Heads of units with around 15 years of experience working with advanced analytics. The lowest amount of experience was four (4) years.

In the first round of interviews, it became clear that many participants were unaware of the changes happening in the algorithm landscape, such as the planned AI legislation from multiple countries. Most of them were aware of the unintended consequences caused by algorithms and could relate to the difficulties in designing and maintaining algorithms' performance during their lifecycle. None of the participants had any review process to evaluate the possibility of unintended consequences; instead, they referred to "continuous improvement," many of the participants realized they were unconsciously utilizing the principles presented in the matrix. The main idea is that the more complex the environment, the more resources are required to design and develop suitable algorithms for that operational environment.

Participants criticized the framework's simplicity and "naivety" toward real-life applications. They mentioned that while the framework makes perfect academic sense, it fails to capture many nuances of real-world scenarios.

The titles of the axes received mixed feedback. Some participants found them clever, while others felt they were too similar and should be modified. The transition areas should be more clearly defined, especially if there are to be any legal requirements for more complex algorithms. It should indicate where the definition and measurement for increased safety measures should be. Additionally, there was a critique that even a simple algorithm can cause devastating effects for individuals in cases like social security benefits decision-making, loan approvals, or judicial decisions where a person might face indictment if the algorithm recommends it. Some participants also noted that the matrix is complex to understand. The matrix emphasizes the environment in which algorithms operate, which was not previously the case. Developers typically focus on the problem they are trying to solve, but this matrix highlights the importance of considering the environment and its limitations.

The positive feedback was that such tools are necessary and represent a way to increase trustworthiness. All interviewees agreed that the trend of increased algorithm involvement will continue and should be carefully regulated.

The suggestions and critiques provided by the participants will be considered in the modified matrix.

Table 9. Interview round one feedback.

Element	Overall feedback from the first interview	Overall feedback from the second interview
Understandability	It wasn't easy to understand. This could have been the failure of communication from the presenter.	
Usefulness	The matrix was seen as too underdeveloped and unable to grasp the infinite nuances of real-world problems.	
Trustworthy	This was difficult to answer; maybe it was the most common answer.	
Future requirements	Definitely.	

4.3 First Iteration on the Algorithm Power Matrix

Based on the interviews, we implemented several enhancements to the framework. Firstly, we refined the categorization within the framework of four quadrants. Secondly, an essential dimension in the matrix is the division of the four quadrants based on the low and high requirements for safety measures. The lowest requirements are in the bottom left corner (Q1), where the number of possible interactions is extremely low, and the control over the algorithm is exceptionally high. This means there is a low possibility for widespread impact since there is a high level of control over the environment in which the algorithm operates. The highest requirement quadrant is the top right (Q4),

where the algorithm has room to maneuver and is more challenging to control due to its scale of interactions. A small team of developers can hardly control the individual interactions of social media algorithms with millions of users or manage every second of a self-driving car and the thousands of decisions its algorithms make. The defining elements can be categorized as:

1. High number of interactions with a high number of other agents and players.
2. Speed of these interactions. No human control has time to intervene in this quadrant.

The two other quadrants lie between these two extremes. The top left quadrant (Q3) represents a semi-open environment with a medium number of interactions and a small chance for human intervention. The bottom right quadrant (Q2) also has a high number of interactions but with a medium chance of control for the developers.
The threshold for moving between quadrants is defined by the possibility of intervening in the algorithm's real-time decision-making. The further this possibility extends, the higher the quadrant the algorithm belongs to. This should be examined more closely in future research.

Key factors are the number of interactions (x-axis) and the level of control developers/algorithms have over their actions in that environment. For example, an algorithm operating within an organization's intranet analyzing site traffic works in a controlled environment.

Finally, the framework's visual aspect was improved to make it easier to understand and grasp the essential idea.

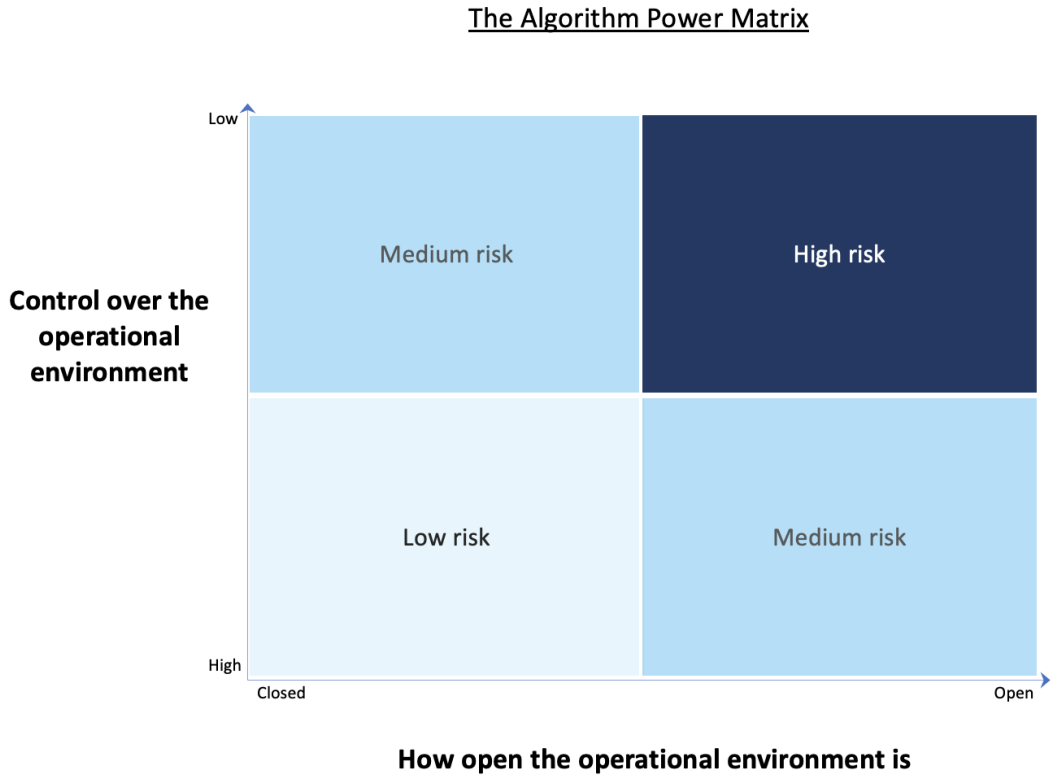

Figure 12. Improved Algorithm Power Matrix framework.

4.4 Second Round of Expert Interviews: Evaluating the Iterated Algorithm Power Matrix

For the second interview, our expert participants remained largely the same. One participant from the first round couldn't participate in the next round, making the total number of participants eight (8).

The second interview provided evidence that communicating the framework's benefits was challenging. However, a greater number of interviewees understood the framework's purpose and potential applications. There was a consensus that the framework

could serve as a guideline and as an additional framework for considering algorithms' lifecycles.

Criticism was again directed at the framework's simplicity, suggesting that it is too general-purpose and may not offer significant value for algorithm development. Instead, it appears more beneficial at the management level, providing managers with tools to categorize their algorithms.

An interesting observation was that individuals with a developer-oriented background did not perceive the need for caution in managing algorithms as strongly as those without a developer background. One interviewee highlighted that the matrix is particularly useful when considering the actual impact of the algorithm on end users. For example, an individual who is denied a job offer or a loan based on algorithmic decisions would be directly affected. This difference in perspective can be attributed to the developers' focus on problem-solving, where the primary goal is to optimize processes. This is an interesting point and would require further evaluation.

Another interesting observation made by one participant was that the matrix might help the developers better define the limits and intended target of their algorithm. This also means that they can reduce the algorithm's complexity and make it easier to operate. Simpler algorithms are usually better since you have a better understanding of their functionality.

Table 10. Interview round two feedback.

Element	Overall feedback from the first interview	Overall feedback from the second interview
Understandability	It wasn't easy to understand. This could have been the failure of communication from the presenter.	The new visuals made it significantly simpler and easier on the eye. Understandability improved, especially for the non-developer-oriented audience.
Usefulness	The matrix was seen as too underdeveloped and unable to grasp the infinite nuances of real-world problems.	There was a broader understanding of use cases for this kind of tool. Especially in managing the algorithm portfolio in higher management.
Trustworthy	This wasn't easy to answer; maybe it was the most common answer.	The answers varied again, and it was concluded that the increase in trustworthiness would need to be adequately tested with a complete development lifecycle in multiple environments and types of algorithms.
Future requirements	Definitely.	Again, the consensus was that this kind of tool would be required in the future, but with significant modifications to make it more user-friendly and to define its shortcomings, such as thresholds for different categories.

4.5 Final Version: Enhancements and Future Directions for the Algorithm Power Matrix

Based on the feedback, we have constructed a final version of the framework.
The following items have been modified or expanded:

Thresholds. The participants clearly needed to understand when their algorithms would enter a different quadrant.

The thresholds should be considered case-by-case, as the requirements differ significantly. The main new contribution regarding the thresholds is that the horizontal thresholds (the x-axis measuring control over the environment) should be tied to the company's available resources. This means that if the company has ample resources to monitor and, if necessary, intervene in the algorithm, it can be labeled as a low-risk algorithm. If there is a mismatch between the algorithm's operating environment requirements and the available resources, it should be categorized as a medium or high-risk algorithm.

The Lowest risk quadrant (Q1) is the area where the company can easily manage the algorithm with its current resources. The following quadrants, Q2 and Q3, are those where the company does not have complete visibility or time to intervene in the algorithm's decision-making process. These quadrants are more complex to manage and require an increased amount of dedicated resources. In these quadrants, mistakes happen but are reversible in a timely manner. In the final quadrant (Q4), the company recognizes that it lacks adequate resources to properly manage the algorithms. This means there is an increased chance of significant incidents with irreversible outcomes.

The thresholds for moving vertically along the y-axis, which measures the openness of the environment, proved too complex for this thesis's scope and will be left for further studies. Instead, we will introduce a table of certain parameters that can inspire further determining the most suitable way of measuring the openness of algorithmic environments. The green color indicates that the element is geared towards a more stable

environment, which decreases the possibility of unwanted consequences. This places the algorithm more towards the low-risk quadrants in the Algorithm Power matrix. The opposite applies to elements marked with red color. The table is based on the work of several authors.

Table 11. Elements to consider when measuring the openness of the algorithm environment.

Element	Explanation	Sources
State of the environment	Are we talking about ***a static environment*** such as a chess board or ***a dynamic environment*** that is in constant change	Su, Huang, Adams, Chang, Beling (2022) Kong, Zhou, Du, Zhou, Zhao (2023) Y. Wang, H. He and C. Sun (2018) Fu, Gao, Liu, Yang, and Zhu (2023)
Pace of change	If the environment is dynamic, then how fast is the change occurring? Are we talking about ***rapid and unknown*** changes or ***mild and known changes***	Kong et al., (2023) Wang et al., (2018) Gao et al., (2023) Kim & Yang (2022)
Observability	What portion of the environment can be ***fully observed*** (e.g., the state of the system is fully known each time), and what portion remains hidden or ***partially observed,*** e.g., the hidden iceberg vs. chessboard	Zhang and Zhao (2024) Kong et al., (2023) Gao et al., (2023)
Age and available data	Is there an abundant amount of ***historical data,*** or is the environment ***completely new*** without possessing the relevant historical data and benchmarks	Gao et al., (2023) Kim & Yang (2022)
Predictability	Can we fully expect to understand the outcome of the algorithm's subsequent actions (e.g., ***Deterministic***), or is there a set of random	Jin, Scheinberg, Xie (2024) Kong et al., (2023) Kim & Yang (2022)

	behaviors creating ***stochastic*** conditions	
The number of possible changes	Do we operate in a rule-based environment with a finite (***discrete***) number of rules, or do we have infinite (***continuous***) amount of possibilities	Kong et al., (2023) Wang et al., (2018) Kim & Yang (2022)
Other players	Do we have to interact with others (***multi-agent***), or are we the only one operational in that environment (***single agent***)	Su et al., (2022) Kong et al., (2023) Gao et al., (2023) Kim & Yang (2022)
Type of players	Should we all ***collaborate,*** or are we in an environment that encourages ***competing*** with each other's	Su et al., (2022) Kong et al., (2023) Gao et al., (2023)
Feedback loop	Do we change our behavior according to the environment, and does the environment change behavior based on our actions? Is there ***adaptive behavior between the algorithm and environment or non-adaptive?*** *Note that this kind of action can be both good and bad*	Su et al., (2022) Kong et al., (2023)

This table offers a simplified list of elements required to assess the openness of an algorithmic environment. As explained by Kong et al., (2023), complexity increases when combining these elements, for example, in a Zero-Sum Semi-Markov Game (ZSSMG), which would include a combination of at least six out of nine elements presented in the table, thus significantly complicating the threshold calculations.

Another framework feature was constructed based on the interview comments about using the matrix as a "portfolio management tool." This means that a matrix can be used as a baseline to visualize the algorithm portfolio, not just individual algorithms. This would allow management to have increased transparency regarding the algorithms in

operation. It can also be used to determine if the requirements of algorithms in the development phase should be modified to lower their risk profile.

The final version of the deliverable, seen in Figure 13 visualizes an example of how an algorithm portfolio would look.

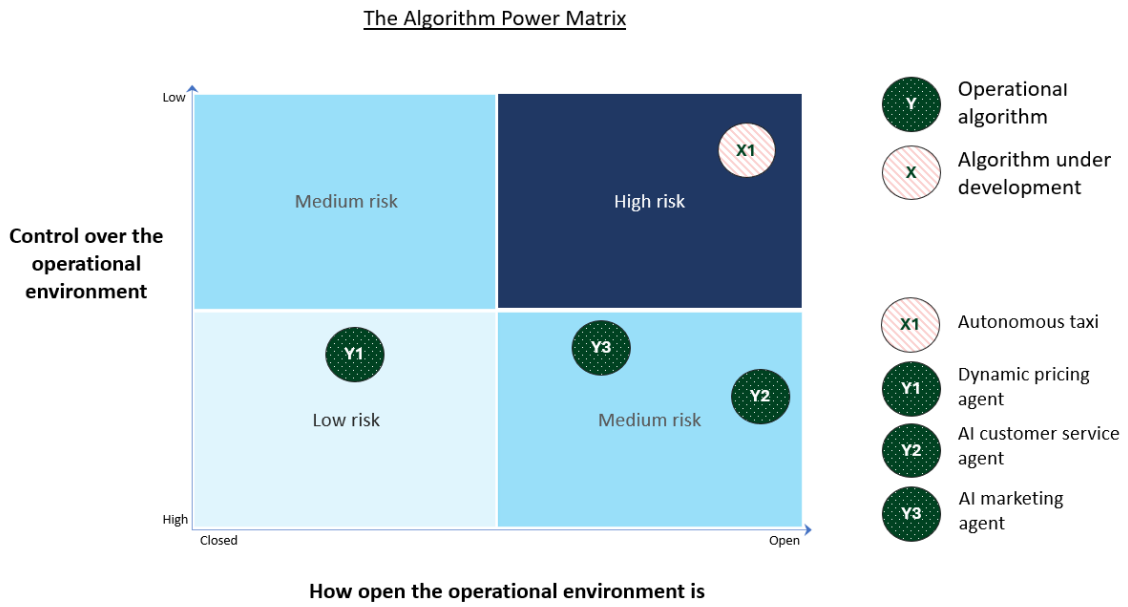

Figure 13. The final version of Algorithm Power Matrix.

In Figure 13, the threshold visualization is incomplete due to the previously mentioned complexities regarding measurement. The matrix visualizes the portfolio view of algorithms; algorithms with a higher risk for unintended consequences are illustrated in red. The algorithms with a lower risk for unintended consequences are illustrated in yellow. This visualization creates a way to facilitate a strategic level discussion by offering easy comparisons between different algorithms, having conversations on resource allocation ethical concerns, and ensuring adequate explainability for highlighted algorithms. An example would be to first classify algorithm y1, which is currently under development, as a high-risk algorithm which would require an additional governance overview to go

through risk evaluation for downgrading the risk category based on the parameters offered in table 11. Table 11 should be utilized for evaluating the transfer of algorithms from one category to another, such as if the algorithm x1 would receive an update that would increase its autonomy and reach over the environment where it operates its risk level should be re-evaluated on the governance level. Governance level re-evaluation of the algorithms should be a regularly reoccurring activity to assess the risks related for the organization algorithm portfolio, comparable for example a financial portfolio where regular adjustments are performed to find optimal and tolerable risk levels. Additionally, table 11 contributes to these discussions by offering elements to consider for ensuring accurate classification of the algorithms and identifying gaps that require additional attention. The algorithm portfolio still needs further evaluation and development, but it provides an interesting avenue for further studies.

In the following chapters, we will provide further links to previous research.

4.5.1 Regulatory Implications to the Algorithm Power Matrix

To link our contribution towards the regulatory concerns, we propose that the tools we created, the Algorithm Power matrix and the Portfolio tool, be used as a toolset to measure and classify algorithms according to the definitions provided by the regulatory forces. Especially as the EU AI Act has proposed categorizing algorithms based on their risk profile, our tools would offer an avenue for reaching a common agreement and understanding of the tools and ways to classify algorithms effectively. This approach aligns well with previous research by Finocchiaro (2024) and Mariani & Dwivedi (2024), which aims to maintain the pace of AI innovation without overflowing AI development with bureaucracy. The Algorithm Power Matrix complements the EU AI Act regulations where algorithm risk is defined as minimal, specific transparency, high, and unacceptable risk levels. The Q4 in our framework should receive special attention for regulatory purposes to examine if the threshold overlaps the unacceptable risk level. This would also help future studies define the appropriate risk profiles for our Power Matrix and Algorithm Portfolio.

We recognize the limitations and the required legal framework to support such operations. Additionally, further studies are needed to specify algorithms' legal requirements, boundaries, and responsibilities.

4.5.2 Integrating Our Findings with Previous Research, Existing Models, and IS Knowledge Framework

Based on the findings, we propose several areas where our research can contribute during the algorithm lifecycle. We review previously presented models and enhance them with our findings. Further research is required to determine the suitability of this preliminary suggestion. We introduce our findings in four (4) previously introduced models for managing algorithm design, development, and deployment represented in the following table.

Table 12. Contributions towards previous research and available toolsets.

Related Prior Work	Our Contribution
ADMS framework (Marabelli et al., 2021).	Extending the lifecycle analysis to incorporate environment-specific risk analysis.
Data Science Trajectory (Martinez-Plumed et al., 2021).	Adding the control factor for the environment as well as the impact consideration explorative action.
ML development pipeline (Eken et al., 2024).	Two additions are a portfolio tool for management purposes and an unintended consequence risk analysis, which require new roles and responsibilities.
The hourglass model (Mäntymäki et al., 2023).	Two additions: portfolio tool for environment-specific governance and risk analysis tool.
Conclusion	Our findings fit well with previous research and offer a complementary set of ways to enhance trustworthiness in several areas. We can also see that

	this kind of perspective was partially lacking in previous research and toolsets.

To highlight our framework's suitability for complementing existing research, we added essential parts of our framework to Marabellis et al., (2021) ADMS framework for managing strategic choices. Essential changes have been highlighted in green and include the environmental factor on where and with what freedom the algorithm would operate.

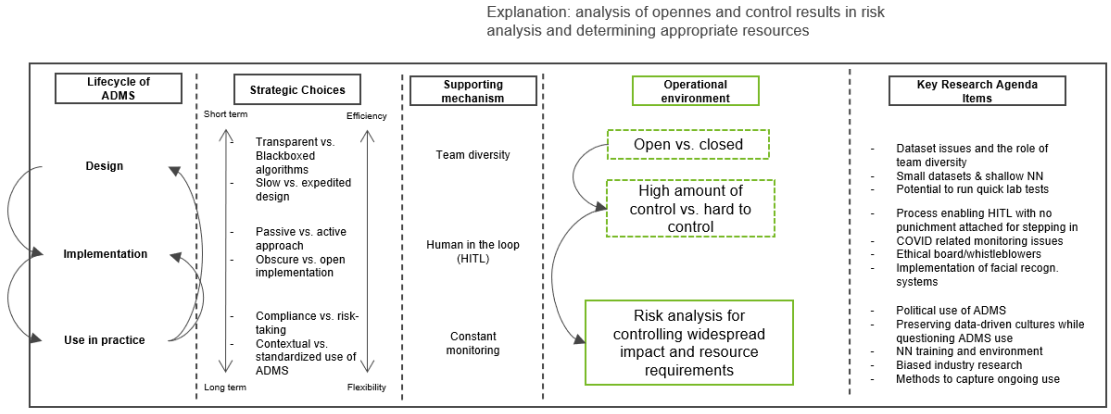

Figure 14. Improved ADMS framework (Marabelli et al., 2021).

Additionally, we have introduced our additions for the Data Science projector workflow from Martinez-Plumed et al., (2021). Again, essential additions are highlighted with green colour. Our contributions again relate to including the environment where and with what freedom the algorithm would operate, with additional factors on analyzing and identifying the potential unintended consequences.

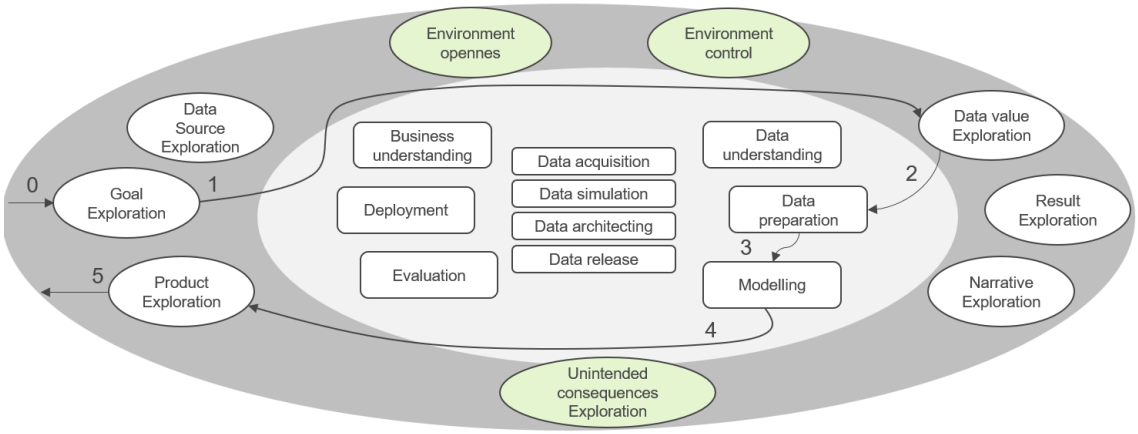

Figure 15. Improved DS trajectory (Martinez-Plumed et al., 2021).

As per the pipeline for ML development and operations by Eken et al., (2024), our findings contribute to two areas. Firstly, the algorithm portfolio management tool complements pipeline management by adding a unified layer to identify the total risk of each algorithm in pipeline management. It is recommended that this tool be accompanied by a new role description for algorithm portfolio managers to ensure a holistic view of all algorithms within their portfolios. Secondly, the algorithm power matrix complements the monitoring capabilities by identifying unintended consequences. The two complementary improvements should be tied to the central piece of this pipeline development, the versioning. Each iteration should provide an additional view for the development team to improve their design to lower the risk of unintended consequences and to improve the trustworthiness of the algorithm in development.

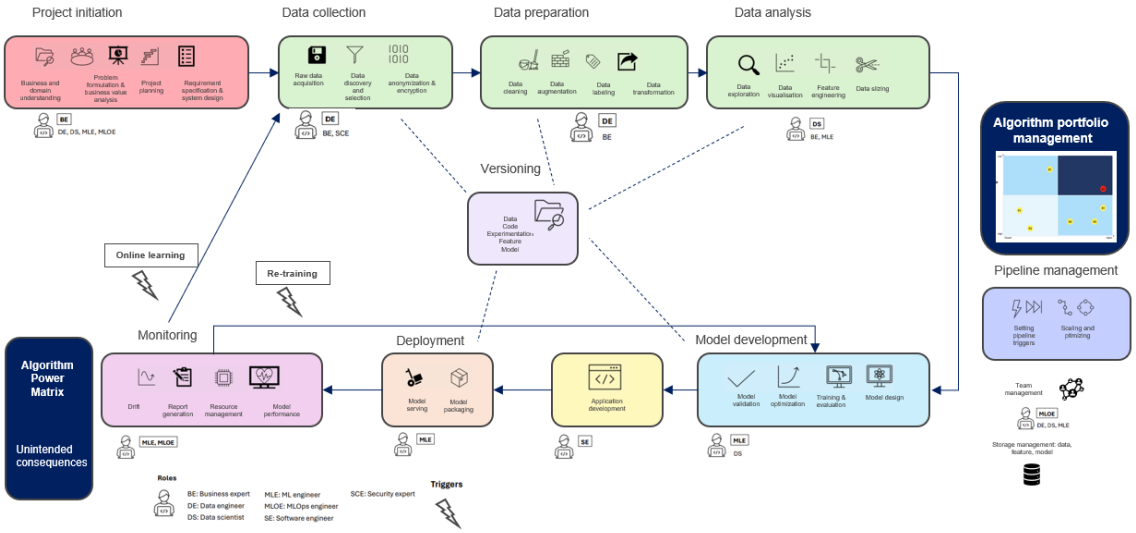

Figure 16. Enhanced ML development pipeline (Eken et al., 2024).

We believe that our findings also provide complementary features for the earlier introduced hourglass model by Mäntymäki et al., (2023). Our findings enhance the hourglass model by installing the algorithm portfolio tool on the organisational layer. This would increase the transparency and visibility of the organisation's existing and future algorithms. As Mäntymäki et al., (2023) explained, this organisational layer is to provide a layer to ensure a strategic alignment within the organisation. The portfolio tool provides one way to ensure that the whole organisation is aligned and that the algorithm development can be viewed from a strategic perspective. The second contribution would be adding the algorithm power matrix to the AI system level, next to the “risk and impact management”. We argue that this would increase the probability of having a family of increasingly trustworthy algorithms as performance and risk monitoring would be added under the portfolio management tool to ensure high transparency.

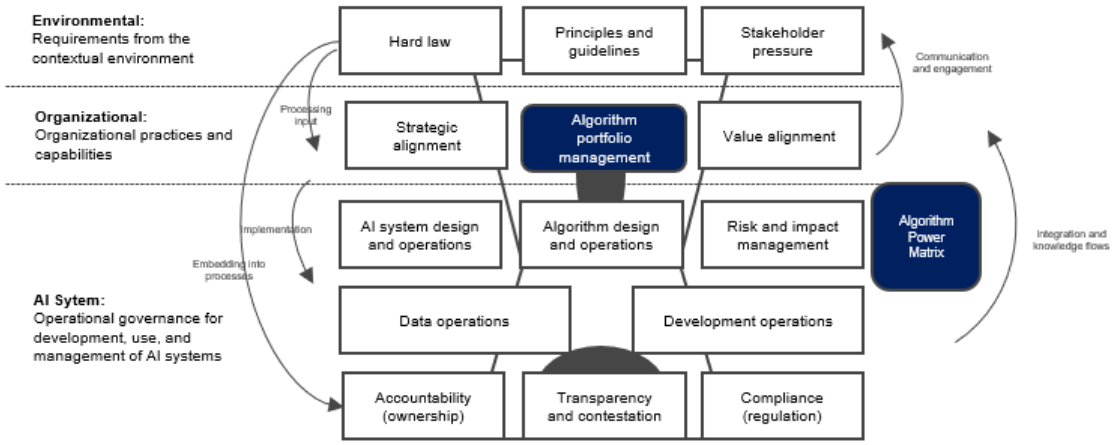

Figure 17. Enhanced hourglass model (Mäntymäki et al., 2023).

Further studies are required to determine the specific working methods and best practices for utilising the tools provided in the examples. For example, it should be determined whether the portfolio should be allocated for all algorithms or grouped by systems based on their intended function, business unit, or some other way, and the ideal risk tolerance level.

Next, we aim to formalise our contributions to the IS research field by utilising the previously presented IS research framework.

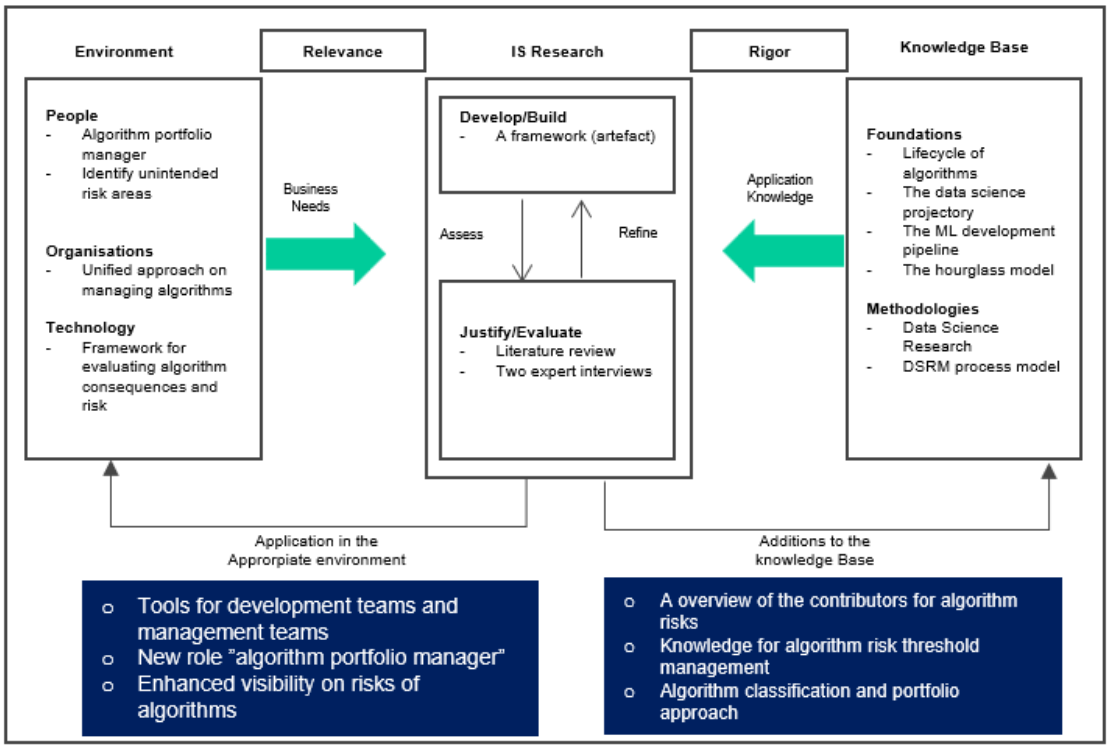

Figure 18. Our contributions to IS research (Hevner et al., 2004).

We evaluate our contribution towards the IS research field by examining the utility and truth of our artefact, as Sein et al., (2011) proposed. The truth comes from prior knowledge, and we have done excessive research on previous literature on multiple domains concerning this matter. We have found many pieces of the puzzle and provided a more comprehensive picture by bringing prior knowledge contributions together. Thus, we have provided an overview of the issue by combining previous research in a novel way. The utility comes from being useful in a business environment, which we ensured by iterating our artefact with experienced industry experts. Following the second interview, we received overall positive feedback from our expert panel, and the utility was confirmed on a slightly different use case than what we first had in mind.

Based on the IS research figure by Hevner et al., (2004), we evaluate our artefact contributions towards the appropriate environment and knowledge base. In our contribution to the knowledge base, we have added an increased understanding of algorithm consequences by discovering numerous instances that need to be considered while

developing trustworthy algorithms. Such as the thresholds for measuring algorithm interactions in an open and closed environment; our summary of the considerations is presented in Table 11. As our study included several other models and frameworks, we have added to the knowledge base by evaluating these prior models based on our findings and discovered areas to improve in the safety considerations. On the other side, we have suggested improvements considering people, organisation and technology. People-related improvements are adding new responsibilities or roles to the algorithm teams, which would require more consideration of the unintended consequences and the algorithm portfolio view. This consideration aligns with the ML pipeline presented by Eken et al., (2024), where we enhanced the original model by introducing a role for algorithm portfolio managers, as seen in Figure 16. As for the organisation, we have identified the need for improved transparency regarding the combined algorithm portfolio impact and consequences. Thus, as presented in the enhanced hourglass model by Mäntymäki et al., (2023), we added an additional layer for organisations to increase precisely this ability. The technology part consists of our Algorithm power matrix, which provides a complementary toolset for evaluating and managing algorithms.

Based on these additions we conclude that there has been a contribution to the IS research on multiple areas which still require further research. Our contribution to the field offers an additional view on algorithm development and management towards trustworthy behavior.

4.5.3 Concluding Previous Chapters: Bringing it All Together

To provide the reader a cohesive reading experience we will utilize this chapter to conclude the learnings of previous chapters and build a case study on how the Algorithm Power Matrix can be applied to a real-world scenario. We analyze two example algorithms previously introduced in this thesis: the EMRAC space exploration algorithm presented by Montanaro et al., (2023) and a genetic healthcare algorithm introduced by Lämmerman et al., (2024). These two algorithms will be examined through three core

concepts: AI regulations, the Algorithm Power Matrix, and the threshold examinations presented in Table 11.

We begin by referencing Table 11 to assess the openness of our chosen algorithms respective operational environments**.** This analysis will later help us to determine their positioning in the Algorithm Power Matrix**,** which will, in turn, link the classification of our chosen algorithms to appropriate regulatory requirements**.** This structured approach demonstrates how our framework can support decision-making in multiple areas related to algorithms and AI.

Table 13. Example case for measuring the openness of algorithms environment.

Element	Algorithm	Explanation
State of the environment	EMRAC	Semi-dynamic environment, operates in Mars and in space exploration.
	Healthcare AI	Dynamic, status of the patients or health professionals can change rapidly.
Pace of change	EMRAC	Environment changes include asteroids and solar flares which can be observed.
	Healthcare AI	Deterioration of patient health can be rapid and unexpected. Amount of patients can change as new patients enter and old leave.
Observability	EMRAC	Space is extremely observable and preparing for changes is good.
	Healthcare AI	Observing patient health relies on constant monitoring which can still fail or cause un-forecasted events. Also, health data can be spread out on multiple databases.
Age and available data	EMRAC	Space environment offers rich and full historical data set.
	Healthcare AI	Trends such as influence seasons are known but on personal level health data can be lacking and un-known.

Predictability	EMRAC	Behavior is largely deterministic with minor exceptions.
	Healthcare AI	A combination on both deterministic and stochastic properties exists such as the un-known side-effects of certain medications or treatments.
The number of possible changes	EMRAC	Space is ruled by the laws of physics, making it discrete environment.
	Healthcare AI	Somewhat discrete but also continuous environment since multiple changes are a possibility.
Other players	EMRAC	Limited amount f other players exists, we can classify it as single-agent environment.
	Healthcare AI	Both single and multi-agent environment are possibility.
Type of players	EMRAC	Collaboration between agents is at least at the moment the dominant trend.
	Healthcare AI	Both are a possibility.
Feedback loop	EMRAC	There is a feedback loop for example on the robot adapting to the space trajectory.
	Healthcare AI	Two-way feedback loop, both the algorithm and doctor or patient can alter their behavior based on the algorithms actions. Making it an adaptive environment.

Based on this classification we can add the EMRAC and Healthcare AI to the Algorithm Power Matrix. We conclude that the EMRAC is operating in closed and controlled environment since its actions doesn't affect other players in the environment and there is a high level of human oversight for space exploring robots due to the organisation structure of space exploration mission control. This aligns well with our previous finding on the thresholds and how the resources of the organisation affect the categorisation of the algorithm. The healthcare AI on the other hand, is operating in an open and dynamic environment with multiple other players such as patients, doctors, and administrations.

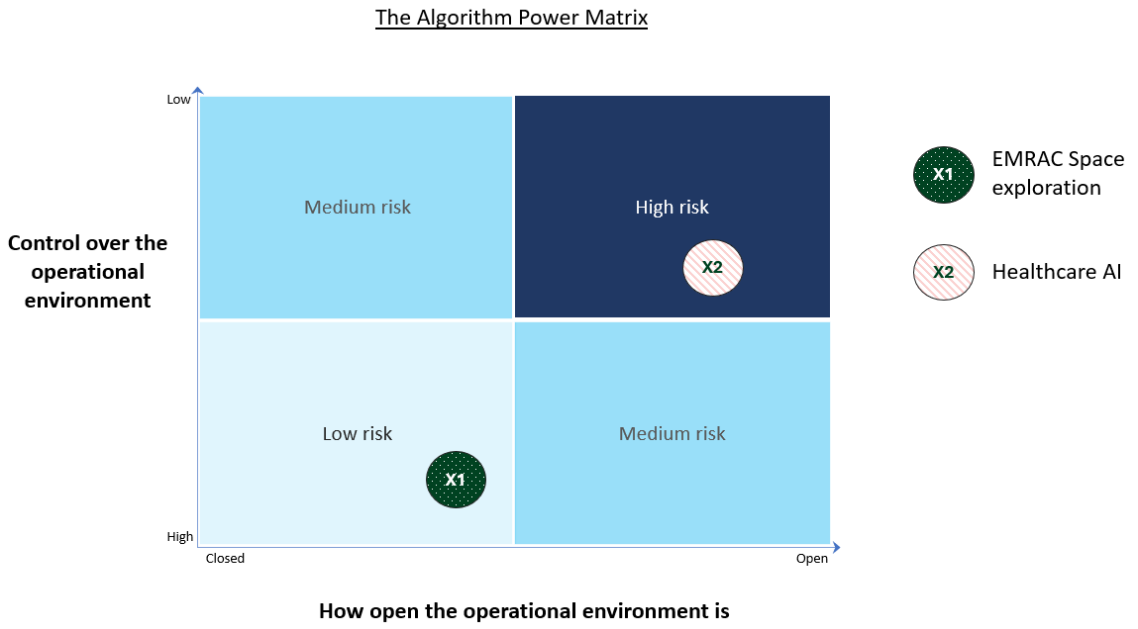

Figure 19. The Algorithm Power Matrix with examples.

As we can see in the Algorithm Power Matrix, the EMRAC algorithm falls into the Quadrant 1**,** classifying it as low risk algorithm since it operates in a strictly controlled environment and has minimal chances for unintended consequences**.** On the other hand, the Healthcare AI system falls into the Quadrant 4**,** categorizing it as high risk due to its potential for direct unintended consequences on human health, ethical concerns, and regulatory violations**.** From this categorisation we can continue to determine the appropriate regulatory measures for these algorithms. The EMRAC algorithm used in a spacecraft is not in a position to cause a major unintended negative consequence so no heavy regulations should be applied to it. The Healthcare AI has the potential to cause unintended negative consequences which can include data leaks or harm to human health. Concluding that this kind of algorithm should be under heavy regulations and strong oversight.

This example illustrates how the Algorithm Power Matrix provides a structured way for evaluating both algorithms and AI systems in a different operational environments. The matrix offers a way for mapping these algorithmic systems and helps equally both

policymakers and industry leaders to integrate appropriate regulatory and governance measures towards algorithms without needlessly delaying innovation and adaption in the field.

5 Discussion

The objective of this study was to contribute to the subject of algorithm trustworthiness by addressing our three research questions:

RQ1: *In what ways can we identify the impacts of different algorithms?*
RQ2: *Can we measure the trustworthiness of algorithms across multiple environments?*
RQ3: *Which factors affect the scale of trustworthiness of different algorithms?*

Our additional objectives for this thesis were:

1. *To examine the broad elements in algorithm development that influence design principles.*
2. *To evaluate the algorithms in diverse environments and contribute towards universal algorithm system design principles that are sector-agnostic.*
3. *To develop frameworks and tools that enable managers to oversee the development, deployment, and operation of algorithms effectively, ensuring alignment with organisational goals and standards.*

These objectives were pursued using Design Science Research as suggested by Peffer et al., (2007), which involves literature analysis and practical iterations with industry experts. Given that the research would result in IS artefacts, we also utilised the research framework from Sein et al., (2011) to ensure both practical applications and theoretical knowledge creation. Our findings indicate that there is a growing need for creating safeguards to ensure trustworthy algorithms. This conclusion is consistent with previous research conducted by Mikalef et al., (2022), Freeman et al., (2022), Balesni et al., (2024), and Meinke et al., (2024).

We began by conducting a comprehensive literature review to provide an extensive overview of algorithm development, introducing various models, frameworks, and a

spectrum of algorithms. We established the connection between human-computer interaction and examined how this interaction can be influenced from both the human and algorithmic perspectives, concluding that both parties are capable of deceptive behavior. Additionally, we provided an overview of regulatory actions taken to ensure the trustworthy development of algorithms and highlighted concerns regarding the over-regulation of algorithm development, which could potentially hinder the pace of innovation.

Based on the literature, an artefact was created, "*The Algorithm Power Matrix*". This artefact was designed to understand the amount of influence the algorithm itself has in any given environment, providing an answer to RQ1. The determination of influence exercised by the algorithm was based on two essential factors: the control of the algorithm by the developer team and the interactions the algorithm has with others in the given environment. The operating environment was chosen as a starting point since it was evident that the trustworthiness of the algorithm depends on its action and actions are performed in certain environments which can be used for examining algorithm actions. Any algorithm that developed unintended consequences couldn't be trustworthy. Therefore, an algorithm that consistently achieves intended outcomes in a given environment is deemed trustworthy.

The environment causes significant challenges for the developers and for the algorithm since the more complex the environment is, the harder it is for the developers to prepare for unintended consequences. For this reason, the control was chosen as one component of the artefact since the greater control allows the developers to mitigate and adjust the impact and behavior of the algorithm. The number of interactions was chosen since the more interactions there are, the harder it is to control the algorithm, thereby raising the possibility of unintended consequences. This was the conclusion based on the literature study and answering RQ2 and RQ3.

The second part of the artefact development was to expose our artefact to a group of experts. This step proved to be vital since without the expert's opinion the matrix wouldn't have been exposed and all the faults in it would have not been found if it was based solely on the literature review.

The original goal was achieved by the creation of the artefact, and it is now clear that the matrix can offer benefits in two ways. Firstly, it can be used as a visualisation tool to place existing algorithms and present them in a meaningful way for the upper management. This would offer them an understanding of the algorithm landscape in which they are currently operating. This is an easy way to visualise the possible risk areas and the algorithms behind them. This would also offer a way for the management to step in and audit the algorithms with the highest risks or establish a governance board to monitor their behavior.

Secondly, this is a way for the developing teams to increase their precision and narrow down the areas where they intend to place their algorithms by combining the number of interactions and control for measuring the risk levels associated with the algorithm. Using the matrix for this would offer them a chance to reduce the associated risk for the algorithm in the first place. Simplifying algorithms is generally considered a good development practice. This matrix allows developers to conduct a reality check and adjust their ambitions to match the desired risk category.

Additionally, we provided a comprehensive table (Table 11) of factors affecting the measurement of the openness of the algorithmic environment. This contributes towards the RQ2 and RQ3 and provides a promising direction for follow-up research.

At the end of our research, we combined our findings with the previous work from multiple authors whose work we had studied to thoroughly understand the algorithm landscape. Most notably, we managed to integrate our findings into four previous models and frameworks which have been actively used through the years for data science and

algorithm development. The first integration was with the ADMS framework from Marabelli et al., (2021), where we added the consideration for environment-specific risk analysis. Our second integration was with the Data Science Trajectory from Martinez-Plumed et al., (2021), where we enhanced the original model by adding our findings on the control of the environment. The third integration was towards the ML development pipeline by Eken et al., (2024), which received enhancement from our findings about algorithm portfolio management and monitoring for any unintended consequences. The fourth integration was more towards the governance level in the hourglass model by Mäntymäki et al., (2023), where we integrated our findings regarding environmental considerations and risk analysis. Additionally, our findings can provide an additional tool for the regulatory authorities for measuring and monitoring the algorithm risk categories effectively without causing unnecessary delays for AI innovation and development as mentioned by Finocchiaro (2024) and Mariani & Dwivedi (2024). Finally, we added a case study on how our findings are applicable for a real-world scenario by introducing two algorithms and their positioning in our findings. Overall, our research provided multiple contributions and further research avenues. Most significantly we contributed to the prior models on algorithm development by introducing the operational environment thinking into these models.

5.1 Recommendations for researchers & practitioners

For practitioners, the usefulness of the guidelines presented here is questionable, at least for the near future. There exists the need to turn the matrix threshold measurement into reality, which was pointed out during the expert interviews but proved to be too complex a topic for this thesis and requires separate and dedicated research activities. Meanwhile, we recommend and encourage the practitioners to take the ideas provided in this study and actively develop them in their own field of expertise. As we pointed out, the tools we developed in this study offer a valuable way of visualising, managing, and communicating the state of algorithm development inside an

organisation. While the research on this topic keeps expanding, there are still no incentives for private companies to start actively sharing information about how their algorithms work. There is always the problem that algorithms are part of companies' property and Intellectual Property Rights (IPR), which doesn't make it easy to give away trade secrets to achieve greater trustworthiness through transparency and explainability. This is why I believe there will be a need in the future for independent auditors and certifications for auditing algorithms, which are targeted especially in the consumer areas. This can also create a business opportunity when the regulation provides increased incentives for private companies to include auditing measures for algorithms, and there are already early signs of new industries being created for this purpose, as noted by Meinke et al., (2024).

For researchers, we recommend utilizing our findings as one complementary part for defining future ways of measuring algorithm trustworthiness and creating safety precautions. Another recommendation for researchers is to examine further the idea of defining clear way for measuring the thresholds on the matrix. Additionally, there is an exciting possibility for utilising our matrix to perform case studies across various environments and industries. For researchers interested in policymaking we recommend on investigating the integration of our matrix and other frameworks and tools for defining adaptive regulations for algorithms based on their risk profile. We also offered one additional way to measure and visualize the abstract nature of algorithm classification. This study culminated the previous work by bringing together the identified issues and one possible solution in a simple format.

As conclusion there seems to be a need to develop tools to fully understand the full impact algorithms can have in our lives. Also, there is a need for developing ever more advanced algorithms. In my opinion, the regulation and risk evaluation frameworks should not be an obstacle for the innovations that advanced algorithms can bring to us.

5.2 Limitations and recommendations for future research

The limitation of this study is the small participant group, and that the algorithm power matrix was never included in the full algorithm development cycle. It remains as a theoretical tool until used in a real-world example. Further limitations are the loosely defined thresholds separating the four different quadrants and the tangible link to integrate regulatory guidelines is also not fully established.

Further studies should be conducted, especially on the implementation of categorizing algorithms and measuring the thresholds of various algorithms. As mentioned already, a clear and comprehensive case studies across industry and with various environments to test the matrix is required. The impact algorithms have on our lives should also require further studies. This is in line with previous research, for example, from Lämmerman et al., (2024).

For designing trustworthy algorithms, consider the various environments where these algorithms operate, the interface between humans and AI vs. Black Box algorithms, and how that is linked in the design phase. Also, regarding whom trustworthiness is reflected, does trustworthiness differ from that of a consumer to that of a developer? Mikalef et al., (2022) also previously explained this.

In conclusion, there are still multiple aspects to be clarified when dealing with trust between man and machine to discover our own Nash equilibrium. In my opinion, there will be unintended consequences, but in the long run, we will become inseparable companions. I remain optimistic about the future.

References

Abbasian, M., Khatibi, E., Azimi, I. et al. Foundation metrics for evaluating the effectiveness of healthcare conversations powered by generative AI. npj *Digit. Med*. 7, 82 (2024). https://doi.org/10.1038/s41746-024-01074-z

Abdel-Karim, B.M., Pfeuffer, N. & Hinz, O. Machine learning in information systems - a bibliographic review and open research issues. *Electron Markets* 31, 643–670 (2021). https://doi.org/10.1007/s12525-021-00459-2

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alpaydin, E. (2010). Introduction to Machine Learning.

Amel-Zadeh, A., & Serafeim, G. (2017). Why and How Investors use ESG Information. Evidence from a Global Survey. *Harvard Business School Working Paper*. http://nrs.harvard.edu/urn-3:HUL.InstRepos:30838135

Arrieta, A. B., Díaz-Rodríguezb, N., Del Sera, J., Bennetot, A., Tabikg, S., Barbadoh, A., Garciag, S., Gil-Lopeza, S., Molinag, S., Benjaminsh, R., Chatilaf, R., & Herrerag, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. *Journal of the Association for Information Systems*, 22(2). DOI: 10.17705/1jais.00664 https://aisel.aisnet.org/jais/vol22/iss2/8

Ashktorab, Zahra & Liao, Vera & Dugan, Casey & Johnson, James & Pan, Qian & Zhang, Wei & Kumaravel, Sadhana & Campbell, Murray. (2020). Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proceedings of the ACM on Human-Computer Interaction*. 4. 1-20. DOI 10.1145/3415167. https://doi.org/10.1145/3415167

Atherton, Daniel. (2024-02-02) Incident Number 634. in Atherton, D. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative. Retrieved on August 15, 2024, from incidentdatabase.ai/cite/634.

Atherton, Daniel. (2023-10-07) Incident Number 573. in Atherton, D. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative. Retrieved on August 15, 2024, from incidentdatabase.ai/cite/573.

Bandi A, Adapa PVSR, Kuchi YEVPK. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet*. 2023; 15(8):260. https://doi.org/10.3390/fi15080260

Bambauer J. & Zarsky T., (20218), The Algorithm Game, 94 *Notre Dame L. Rev*. 1. Available at: https://scholarship.law.nd.edu/ndlr/vol94/iss1/1

Balesni, M., Hobbhahn, M., Lindner, D., Meinke, A., Korbak, T., Clymer, J., Shlegeris, B., Scheurer, J., Stix, C., Shah, R., Goldowsky-Dill, N., Braun, D., Chughtai, B., Evans, O., Kokotajlo, D., & Bushnaq, L. (2024). Towards evaluations-based safety cases for AI scheming. arXiv. https://arxiv.org/abs/2411.03336

Batarseh, F.A., Freeman, L. & Huang, CH. A survey on artificial intelligence assurance. J *Big Data* 8, 60 (2021). https://doi.org/10.1186/s40537-021-00445-7

Batool, Amna & Zowghi, Didar & Bano, Muneera. (2024). AI Governance: A Systematic Literature Review. 10.21203/rs.3.rs-4784792/v1. https://doi.org/10.48550/arXiv.2401.10896

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671-732. Available at http://dx.doi.org/10.2139/ssrn.2477899

Baskerville, Richard; Baiyere, Abayomi; Gregor, Shirley; Hevner, Alan; and Rossi, Matti (2018) "Design Science Research Contributions: Finding a Balance between Artefact and Theory," *Journal of the Association for Information Systems*, 19(5),. Available at: https://aisel.aisnet.org/jais/vol19/iss5/3

Beiranvand, V., Hare, W. & Lucet, Y. Best practices for comparing optimization algorithms. *Optim Eng* 18, 815–848 (2017). https://doi.org/10.1007/s11081-017-9366-1

Belanche, D., Casaló, L. V., Flavián, M., & Ibáñez-Sánchez, S. (2021). Building influencers' credibility on Instagram: Effects on followers' attitudes and behavioral responses toward the influencer. *Journal of Retailing and Consumer Services*, 61, 102585. https://doi.org/10.1016/j.jretconser.2021.102585

Ben Schneiderman (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495-504. DOI: 10.1080/10447318.2020.1741118. https://doi.org/10.1080/10447318.2020.1741118

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, 22(2), 281-303. doi: 10.17705/1jais.00662 Available at: https://aisel.aisnet.org/jais/vol22/iss2/10

Bengio Y., et al., Managing extreme AI risks amid rapid progress. *Science 384*,842-845(2024). DOI:10.1126/science.adn0117

Bhatta, T.R. (2021). False Negative/False Positive. In: Gu, D., Dupre, M.E. (eds*) Encyclopedia of Gerontology and Population Aging*. Springer, Cham. https://doi.org/10.1007/978-3-030-22009-9_572

Bhatt, M., Chennabasappa, S., Nikolaidis, C., Wan, S., Evtimov, I., Gabi, D., Song, D., Ahmad, F., Aschermann, C., Fontana, L., Frolov, S., Giri, R. P., Kapil, D., Kozyrakis, Y., LeBlanc, D., Milazzo, J., Straumann, A., Synnaeve, G., Vontimitta, V., Whitman, S., & Saxe, J. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models. arXiv preprint arXiv:2312.04724. https://doi.org/10.48550/arXiv.2312.04724

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*, 0(0). https://doi.org/10.1177/14614448241253138

Bommasani et al. (2021), §1.1, On The Opportunities and Risks of Foundational Models, Authored by the Center for Research on Foundation Models (CRFM) at the *Stanford Institute for Human-Centered Artificial Intelligence (HAI).* https://doi.org/10.48550/arXiv.2108.07258

Bonnefon, Jean-François & Shariff, Azim & Rahwan, Iyad. (2016). The Social Dilemma of Autonomous Vehicles. Science. 352. http://dx.doi.org/10.1126/science.aaf2654

Bonsón, E., Lavorato, D., Lamboglia, R., & Mancini, D. (2021). Artificial intelligence activities and ethical approaches in leading listed companies in the European Union. International *Journal of Accounting Information Systems*, 43, 100535. https://doi.org/10.1016/j.accinf.2021.100535

Bowen, L., & Wu, L. (2021). AI ON DRUGS: CAN ARTIFICIAL INTELLIGENCE ACCELERATE DRUG DEVELOPMENT? EVIDENCE FROM A LARGE-SCALE EXAMINATION OF BIO-PHARMA FIRMS. *MIS Quarterly*, 45(3), 1451-1482. DOI: 10.25300/MISQ/2021/16565

Brams, S. J. and Davis, . Morton D. (2024, July 31). game theory. *Encyclopedia Britannica*. https://www.britannica.com/science/game-theory

Brink, H., Richards, J. W., & Fetherolf, M. (2017). Real-World Machine Learning.

Brkan, M. (2019). The consequences of the right to data portability in the GDPR: A data-driven analysis. *International Data Privacy Law*, 9(3), 181-196. http://dx.doi.org/10.1016/j.clsr.2017.10.003

Brownell, K. D., & Warner, K. E. (2009). The perils of ignoring history: Big Tobacco played dirty and millions died. How similar is Big Food? *Milbank Q,* 87(1), 259-94. doi: 10.1111/j.1468-0009.2009.00555.x https://doi.org/10.1111/j.1468-0009.2009.00555.x

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Bullynck Maarten. Histories of algorithms: Past, present and future. *Historia Mathematica,* 2015, 43 (3), pp.332 - 341. ff10.1016/j.hm.2015.07.002ff. ffhalshs-01215943

Burton, J. W., Stein, M-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *J Behav Dec Making*, 33, 220–239. https://doi.org/10.1002/bdm.2155

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-12.d https://doi.org/10.1177/2053951715622512

Cambridge University Press. (n.d.). *Algorithm*. In *Cambridge Dictionary*. Retrieved December 29, 2024, from https://dictionary.cambridge.org/dictionary/english/algorithm

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730). https://doi.org/10.1145/2783258.2788613, doi 10.1145/2783258.2788613

Cavalli Francesco, (2024) The State Of Deepfakes 2024, retrieved August 15th 2024, from: https://sensity.ai/reports/

Chakraborty, A., Kale, S., McSherry, F., & Raskar, R. (2018). Interpretability of deep learning models: A survey of results. In 2017 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1-6). *IEEE*. https://api.semanticscholar.org/CorpusID:39860496

China’s *Ministry of Science and Technology*, 2021, The "Code of Ethics for the Next Generation of Artificial Intelligence" https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html

Chen, J., Zhang, T., & Wang, X. (2023). Program of Thought Prompting for Automated Reasoning. arXiv preprint arXiv:2310.04572. https://doi.org/10.48550/arXiv.2211.12588

Cheng, X., Su, L., Luo, R., Benitez, J., & Cai, S. (2021). The good, the bad, and the ugly: impact of analytics and artificial intelligence-enabled personal information collection on privacy and participation in ridesharing. *European Journal of Information Systems*. DOI: 10.1080/0960085X.2020.1869508 https://doi.org/10.1080/0960085X.2020.1869508

Cho, J-H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., & Beaumont, M. (2019). STRAM: Measuring the Trustworthiness of Computer-Based Systems. *ACM Computing Surveys,* 51(6). DOI:10.1145/3277666 https://doi.org/10.1145/3277666

Choi, T. R., & Drumwright, M. E. (2021). "OK, Google, why do I use you?" Motivations, post-consumption evaluations, and perceptions of voice AI assistants. *Telematics and Informatics*, 62, 101628. https://doi.org/10.1016/j.tele.2021.101628

Christian Meske, Enrico Bunde, Johannes Schneider & Martin Gersch (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53-63. DOI: 10.1080/10580530.2020.1849465.https://doi.org/10.1080/10580530.2020.1849465

Chu, Z., Zhang, H., & Liu, B. (2023). Navigate through Enigmatic Labyrinth: A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. arXiv preprint arXiv:2309.15402. https://api.semanticscholar.org/CorpusID:263153015

Ciaran Hughes, Joshua Isaacson, Anastasia Perry, Ranbel F. Sun, Jessica Turner (2021) *Quantum Computing for the Quantum Curious*, eBook ISBN 978-3-030-61601-4 Published: 22 March 2021, DOI: https://doi.org/10.1007/978-3-030-61601-4 pages 81-84

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2009). Introduction to Algorithms (3rd ed.). *MIT Press.*

Cox, L. A., Jr. (2021). Information Structures for Causally Explainable Decisions. *Entropy*, 23, 601. https://doi.org/10.3390/e23050601

Cyberspace Administration of China, 2023, The "Interim Measures for the Management of Generative AI Services" https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Danaher, J., Hogan, M. J., Noone, C., et al. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society.* doi:10.1177/2053951717726554 https://doi.org/10.1177/2053951717726554

Dagar, D., Vishwakarma, D.K. A literature review and perspectives in deepfakes: generation, detection, and applications. *Int J Multimed Info Retr* 11, 219–289 (2022). https://doi.org/10.1007/s13735-022-00241-w

Das, S. D., Bala, P. K., & Mishra, A. N. (2023). Towards Defining a Trustworthy Artificial Intelligence System Development Maturity Model. *Journal of Computer Information Systems*, 1–22. https://doi.org/10.1080/08874417.2023.2251443

Dastgerdi, Karimi, Amin & Javdani Gandomani, Taghi. (2021). On the Appropriate Methodologies for Data Science Projects. 667-673. 10.1109/ICIT52682.2021.9491712. https://www.researchgate.net/publication/353484542_On_the_Appropriate_Methodologies_for_Data_Science_Projects

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?. *Philos. Technol.,* 31, 525–541. https://doi.org/10.1007/s13347-017-0293-z

De Silva, D., & Alahakoon, D. (2021). *An Artificial Intelligence Life Cycle: From Conception to Production*. arXiv preprint arXiv:2108.13861. https://doi.org/10.48550/arXiv.2108.13861

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology*: Applied, 22(3), 331–349. https://doi.org/10.1037/xap0000092

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Sci Rep*, 6, 37825. doi: 10.1038/srep37825 https://doi.org/10.1038/srep37825

Demetis, D., & Lee, A. S. (2018). When Humans Using the IT Artefact Becomes IT Using the Human Artefact. *Journal of the Association for Information Systems*, 19(10). Available at: https://aisel.aisnet.org/jais/vol19/iss10/5

Dhanesh G.S, G. Duthler (2019) Relationship management through social media influencers: effects of followers' awareness of paid endorsement *Publ. Relat. Rev*., 45

(3) (2019), 10.1016/j.pubrev.2019.03.002 https://doi.org/10.1016/j.pubrev.2019.03.002

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming Algorithm Aversion: People will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3), 1155-1170. http://dx.doi.org/10.1287/mnsc.2016.2643

Dinev, T., McConnell, A., & Smith, H. J. (2015). Thinking Outside the "APCO" Box. *Information Systems Research,* 26(4), 639–655. http://dx.doi.org/10.1287/isre.2015.0600

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, 3(2). https://doi.org/10.1177/2053951716665128

Duportail et al. (2024, 8th of July) Undress or fail: Instagram's algorithm strong-arms users into showing skin. *AlgorithmWatch*. https://algorithmwatch.org/en/instagram-algorithm-nudity/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, June). Fairness through awareness. In Proceedings of the 3rd *innovations in theoretical computer science conference* (pp. 214-226). https://doi.org/10.48550/arXiv.1104.3913

Eken, B., Pallewatta, S., Tran, N.K., Misirli, A.T., & Babar, M.A. (2024). A Multivocal Review of MLOps Practices, Challenges and Open Issues. *ArXiv, abs/2406.09737*. https://doi.org/10.48550/arXiv.2406.09737

Elkin-Koren N. (2020) Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*. July 2020. doi:10.1177/2053951720932296 https://doi.org/10.1177/2053951720932296

Enrique, B., Domenica, L., Rita, L., & Daniela, M. (2021). Artificial intelligence activities and ethical approaches in leading listed companies in the European Union. International *Journal of Accounting Information Systems*, 43, 100535. https://doi.org/10.1016/j.accinf.2021.100535

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press.

European Commission, Directorate-General for Communications Networks, Content and Technology, *Policy and investment recommendations for trustworthy AI*, Publications Office of the European Union, 2019, https://data.europa.eu/doi/10.2759/465913

EU Directorate-General for Communication, 2024 AI Act enters into force, https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

Ezugwu, Absalom & Ho, Yuh-Shan & Egwuche, Ojonukpe & Ekundayo, Olufisayo & van der Merwe, Annette & Saha, Apu Kumar & Pal, Jayanta. (2024). Classical Machine Learning: Seventy Years of Algorithmic Learning Evolution. *Data Intelligence*. 10.3724/2096-7004.di.2024.0051. https://api.semanticscholar.org/CorpusID:271710529

Fan, D., Breslin, D., Callahan, J. L., & Iszatt-White, M. (2022). Advancing literature review methodology through rigour, generativity, scope and transparency. International *Journal of Management Reviews,* 24(2), 171–180. https://doi.org/10.1111/ijmr.12291 Go

Fernando H. F. Botelho (2021) Accessibility to digital technology: Virtual barriers, real opportunities, *Assistive Technology*, 33:sup1, 27-34, DOI: 10.1080/10400435.2021.1945705 https://doi.org/10.1080/10400435.2021.1945705

Ferratt, T. W., Prasad, J., & Dunne, E. J. (2018). Fast and Slow Processes Underlying Theories of Information Technology Use. *Journal of the Association for Information Systems,* 19(1). DOI: 10.17705/1jais.00482. http://dx.doi.org/10.17705/1jais.00477

Ferragina, P. (2008). String Algorithms and Data Structures. arXiv preprint arXiv:0801.2378. https://doi.org/10.48550/arXiv.0801.2378

Feuerriegel, S., Hartmann, J., Janiesch, C. et al. Generative AI. *Bus Inf Syst Eng* 66, 111–126 (2024). https://doi.org/10.1007/s12599-023-00834-7

Financial Times. (2024). *Who killed the ESG party?* Retrieved 10th of August 2024 from https://www.ft.com/video/1eeebd90-25d4-4421-a175-deedcdbf9c18

Finocchiaro, G. (2024). The regulation of artificial intelligence. *AI & Society*, 39(6), 1961-1968. https://doi.org/10.1007/s00146-023-01650-z

Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau & Langtao Chen (2023) Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, *Journal of Information Technology Case and Application Research*, 25:3, 277-304, DOI:10.1080/15228053.2023.2233814 https://doi.org/10.1080/15228053.2023.2233814

Fournier-Tombs, E. (2021). Towards a United Nations Internal Regulation for Artificial Intelligence. *Big Data & Society*. https://doi.org/10.1177/20539517211039493

Freeman, L., Batarseh, F., Kuhn, R., Raunak, M. S., & Kacker, R. (2022). Widescale adoption of intelligent algorithms requires that Artificial Intelligence (AI) engineers provide assurances that an algorithm will perform as intended. *Computer (IEEE Computer)*, 55(3), 82-86. https://doi.org/10.1109/MC.2021.3129027

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, 45, 1527-1556. DOI: 10.25300/MISQ/2021/16553. https://ssrn.com/abstract=3879937

Garimella, K., Smith, T., Weiss, R., & West, R. (2021). Political Polarization in Online News Consumption. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 152-162. https://doi.org/10.1609/icwsm.v15i1.18049

Garcia et al., 2015, FEATURE-EXTRACTION-BASED IMAGE SCORING, US 8,929,615 B2, US Patent https://patentimages.storage.googleapis.com/b2/77/0b/4fe4ab3a5d2b88/US8929615.pdf

G. Fu, Y. Gao, L. Liu, M. Yang, and X. Zhu, "UAV Mission Path Planning Based on Reinforcement Learning in Dynamic Environment," *Journal of Function Spaces*, vol. 2023, Article ID9708143, 11, 2023. https://doi.org/10.1155/2023/9863415

Gonzalez v. Google LLC, 598 U.S. ___ (2023) No. 2:16-cv-03282 Gonzalez v. Google LLC :: 598 U.S. ___ (2023) :: Justia US Supreme Court Center

Goel, S., Williams, K., & Dincelli, E. (2017). Got Phished? Internet Security and Human Vulnerability. *Journal of the Association for Information Systems*, 18(1). DOI: 10.17705/1jais.00447. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1761&context=jais

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys* (CSUR), 51(5), 1–42. https://doi.org/10.1145/3236009.

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). *Thousands of AI authors on the future of AI*. arXiv preprint arXiv:2401.02843. https://doi.org/10.48550/arXiv.2401.02843

Gregor, Shirley & Hevner, Alan. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*. 37. 337-356. 10.25300/MISQ/2013/37.2.01. https://doi.org/10.25300/MISQ/2013/37.2.01

Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, Lee MJ, Asadi H. (2019) Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. AJR Am J Roentgenol. 2019 Jan;212(1):38-43. doi: 10.2214/AJR.18.20224. Epub 2018 Oct 17. Erratum in: AJR Am J Roentgenol. 2019 Feb;212(2):479. doi: 10.2214/AJR.18.20994. PMID: 30332290. https://doi.org/10.2214/ajr.18.20224

Hasan, R., Shams, R., & Rahman, M. (2021). Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research*, 131, 591-597. https://doi.org/10.1016/j.jbusres.2020.12.012

Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2019). Cooperating with the future. Nature, 571(7764), 45-50. https://doi.org/10.1038/nature13530

Hind Benbya, Stella Pachidi, Sirkka L. Jarvenpaa (2021) Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems* (2021) 22(2), 281-303 doi: 10.17705/1jais.00662 https://doi.org/10.17863/CAM.65841

Herkert, J., Borenstein, J. & Miller, K. The Boeing 737 MAX: Lessons for Engineering Ethics. *Sci Eng Ethics* 26, 2957–2974 (2020). https://doi.org/10.1007/s11948-020-00252-y

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Holland G., M. Tiggemann A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes *Body Image*, 17 (2016), pp. 100-110, 10.1016/j.bodyim.2016.02.008 https://doi.org/10.1016/j.bodyim.2016.02.008

Hope Koch, Wallace Chipidza, Timothy R. Kayworth (2021). Realizing value from shadow analytics: A case study. *The Journal of Strategic Information Systems*, 30(2), 101668. https://doi.org/10.1016/j.jsis.2021.101668.

Horowitz, M. C., & Scharre, P. (2015). Meaningful human control in weapon systems: A primer. *Center for a New American Security*.

Hosseinmardi H, Ghasemian A, Rivera-Lanas M, Horta Ribeiro M, West R, Watts DJ. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proc Natl Acad Sci U S A*. 2024 Feb 20;121(8):e2313377121. doi: 10.1073/pnas.2313377121. Epub 2024 Feb 13. PMID: 38349876; PMCID: PMC10895271. https://doi.org/10.48550/arXiv.2308.10398

Huang H, Tan H, Xu X, Zhang J, Zhao Z. LACE: Low-Cost Access Control Based on Edge Computing for Smart Buildings. *Electronics.* 2023; 12(2):412. https://doi.org/10.3390/electronics12020412

Hu, L., Zhao, X., & Feng, J. (2024). A Theoretical Understanding of Chain-of-Thought. arXiv preprint arXiv:2402.18312. https://doi.org/10.48550/arXiv.2411.11984

Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., & Sharma, M. (2024). Best-of-N Jailbreaking. arXiv preprint arXiv:2412.03556. https://doi.org/10.48550/arXiv.2412.03556

Härkönen, T., Vänskä, R., Vahti, J., & Lehtonen, K. (2022). Digivallan jalanjäljillä, miten datan avulla voidaan vaikuttaa päättäjiin ja ohjata maailmaa. *Sitra*. https://www.sitra.fi/app/uploads/2022/05/sitra-digivallan-jaljilla-v2.pdf.

Høffding, S., Martiny, K. & Roepstorff, A. Can we trust the phenomenological interview? Metaphysical, epistemological, and methodological objections. *Phenom Cogn Sci* 21, 33–51 (2022). https://doi.org/10.1007/s11097-021-09744-z

Iivari, Juhani & Hansen, Magnus & Haj-Bolouri, Amir. (2021). A Proposal for Minimum Reusability Evaluation of Design Principles. *European Journal of Information Systems*. 30. 286-303. 10.1080/0960085X.2020.1793697. https://doi.org/10.1080/0960085X.2020.1793697

Jaigirdar F.T., C. Rudolph, G. Oliver, D. Watts and C. Bain, "What Information is Required for Explainable AI? : A Provenance-based Research Agenda and Future Challenges," 2020 *IEEE 6th International Conference on Collaboration and Internet Computing* (CIC), Atlanta, GA, USA, 2020, pp. 177-183, doi: 10.1109/CIC50333.2020.00030. https://doi.org/10.1109/CIC50333.2020.00030

Jianyu Su, Jing Huang, Stephen Adams, Qing Chang, Peter A. Beling, Deep multi-agent reinforcement learning for multi-level preventive maintenance in manufacturing systems, *Expert Systems with Applications*, (2022) https://doi.org/10.1016/j.eswa.2021.116323.

Jin, B., Scheinberg, K., & Xie, M. (2024). Sample complexity analysis for adaptive optimization algorithms with stochastic oracles. *Math. Program*. https://doi.org/10.1007/s10107-024-02078-z.

Jinghan Zeng (2020) Artificial intelligence and China's authoritarian governance *International Affairs* 96: 6 (2020) 1441–1459; doi: 10.1093/ia/iiaa172 https://api.semanticscholar.org/CorpusID:228880810

Jon Aaen, Jeppe Agger Nielsen & Andrea Carugati (2022). The dark side of data ecosystems: A longitudinal study of the DAMD project. *European Journal of Information Systems*, 31(3), 288-312. DOI: 10.1080/0960085X.2021.1947753. https://doi.org/10.1080/0960085X.2021.1947753

Kankanhalli, Atreyi (2024) "Peer Review in the Age of Generative AI," *Journal of the Association for Information Systems,* 25(1), 76-84. DOI: 10.17705/1jais.00865 Available at: https://aisel.aisnet.org/jais/vol25/iss1/9

Karpus et al.,(2021) Algorithm exploitation: Humans are keen to exploit benevolent AI *iScience* 24, 102679 June 25, 2021 a https://doi.org/10.1016/ j.isci.2021.102679

Keles B., N. McCrae, A. Grealish A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents *International Journal of Adolescence and Youth,* 25 (1) (2020), pp. 79-93, 10.1080/02673843.2019.1590851 https://doi.org/10.1080/02673843.2019.1590851

Khormali,A.;Yuan,J.-S. ADD: Attention-Based DeepFake Detection Approach. *Big Data Cogn. Comput*.2021,5,49. https://doi.org/10.3390/bdcc5040049

Kiyasseh, D., Cohen, A., Jiang, C. et al. A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. *Nat Commun* 15, 1808 (2024). https://doi.org/10.1038/s41467-024-46000-9

Kim,D.-H., López,G.; Kiedanski, D.; Maduako, I.; Ríos, B.; Descoins, A.; Zurutuza, N.; Arora, S.; Fabian, C. Bias in Deep Neural Networks in Land Use Characterization for International Development. *Remote Sens.* 2021, 13, 2908. https://doi.org/10.3390/rs13152908

Kim, J., & Yang, G. H. (2022). Improvement of Dynamic Window Approach Using Reinforcement Learning in Dynamic Environments. *Int. J. Control Autom. Syst*., 20, 2983–2992. https://doi.org/10.1007/s12555-021-0462-9.

Kirsten Martin (2019). Designing Ethical Algorithms. June 2019 (18:2) | *MIS Quarterly* Executive. DOI:10.17705/2msqe.00012. https://aisel.aisnet.org/misqe/vol18/iss2/5?utm_source=aisel.aisnet.org%2Fmisqe%2Fvol18%2Fiss2%2F5&utm_medium=PDF&utm_campaign=PDFCoverPages

K K, Ramachandran. (2024). DATA SCIENCE IN THE 21ST CENTURY: EVOLUTION, CHALLENGES, AND FUTURE DIRECTIONS. 1. 1-01.

https://www.researchgate.net/publication/377598352_DATA_SCIENCE_IN_THE_21ST_CENTURY_EVOLUTION_CHALLENGES_AND_FUTURE_DIRECTIONS

Kong, W., Zhou, D., Du, Y., Zhou, Y., Zhao, Y. (2023). Hierarchical multi-agent reinforcement learning for multi-aircraft close-range air combat. *IET Control Theory Appl*., 17, 1840–1862. https://doi.org/10.1049/cth2.12413.

Kordzadeh Nima & Maryam Ghasemaghaei (2022) Algorithmic bias: review, synthesis, and future research directions, *European Journal of Information Systems*, 31:3, 388-409, DOI: 10.1080/0960085X.2021.1927212 https://doi.org/10.1080/0960085X.2021.1927212

Korpela, J., & Korpela, M. (2021). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?. *Philos. Technol.,* 31, 525–541. https://doi.org/10.1007/s13347-017-0293-z.

Kosinski, M., Celik, E., Kern, M. L., Koukladas, N., & Sen, S. (2023). Theory of Mind in GPT: Limits and Beyond. arXiv preprint arXiv:2307.09288. https://doi.org/10.48550/arXiv.2307.09288

Kour, R., Karim, R., Dersin, P. (2024). Game Theory and Cyber Kill Chain: A Strategic Approach to Cybersecurity. In: Kumar, U., Karim, R., Galar, D., Kour, R. (eds) International Congress and Workshop on Industrial AI and eMaintenance 2023. IAI 2023. *Lecture Notes in Mechanical Engineering. Springer*, Cham. https://doi.org/10.1007/978-3-031-39619-9_33

Krasnova, H., Widjaja, T., Buxmann, P., Wenninger, H., & Benbasat, I. (2015). Why following friends can hurt you: An exploratory investigation of the effects of envy on social networking sites among college-age users. *Information Systems Research*, 26(3), 585–605. https://doi.org/10.1287/isre.2015.0588.

Krishnaraj P. Rita S. Jitendra Jaiswal, 2024, Comparing Machine Learning Techniques for Loan Approval Prediction, *IACIDS, EAI* DOI: 10.4108/eai.23-11-2023.2343174 http://dx.doi.org/10.4108/eai.23-11-2023.2343174

Lai, Vivian and Tan, Chenhao (2019) On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection isbn: 9781450361255}, *Association for Computing Machinery* doi: 10.1145/3287560.3287590 https://doi.org/10.1145/3287560.3287590

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluation AI tools based on experts' know-what. *Management Information Systems Quarterly*, 45(3b), 1501-1525. Available at SSRN: https://ssrn.com/abstract=3839601.

Li, K., Griffin, M. A., Barker, T., Prickett, Z., Hodkiewicz, M. R., Kozman, J., & Chirgwin, P. (2023). Embedding data science innovations in organizations: a new workflow approach. *Data-Centric Engineering*, *4*, e26. doi:10.1017/dce.2023.22 https://doi.org/10.1017/dce.2023.22

Lisa Marie Giermindl, Franz Strich, Oliver Christ, Ulrich Leicht-Deobald & Abdullah Redzepi (2021). The dark sides of people analytics: reviewing the perils for organisations and employees. *European Journal of Information Systems*. DOI: 10.1080/0960085X.2021.1927213. https://doi.org/10.1080/0960085X.2021.1927213

Liu and Goodhue (2012) Two Worlds of Trust. Information Systems Research 23(4), pp. 1246–1262, http://dx.doi.org/10.1287/isre.1120.0424

Liu-Thompkins, Y., Okazaki, S. & Li, H. Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *J. of the Acad. Mark. Sci.* 50, 1198–1218 (2022). https://doi.org/10.1007/s11747-022-00892-5

Liu, Y., Li, W., Wang, L. *et al.* Why greenwashing occurs and what happens afterwards? A systematic literature review and future research agenda. *Environ Sci Pollut Res* 30, 118102–118116 (2023). https://doi.org/10.1007/s11356-023-30571-z

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005.

Luis Lämmermann, Peter Hofmann, Nils Urbach,(2024) Managing artificial intelligence applications in healthcare: Promoting information processing among

stakeholders, *International Journal of Information Management*, Volume 75, 2024, 102728, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2023.102728.

MacDonald, T. W. (2023). “How it actually works”: Algorithmic lore videos as market devices. *New Media & Society*, 25(6), 1412-1431. https://doi.org/10.1177/14614448211021404

Matias, J. N. (2023, May 10). Humans and algorithms work together — so study them together. *Nature*. https://www.nature.com/articles/d41586-023-01521-z

Marco Marabelli, Sue Newell, Valerie Handunge (2021). The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. *The Journal of Strategic Information Systems,* 30(3), 101683. https://doi.org/10.1016/j.jsis.2021.101683.

Mariani, M., & Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research,* 175, 114542. https://doi.org/10.1016/j.jbusres.2024.114542

Mariia Golovianko, Svitlana Gryshko, Vagan Terziyan & Tuure Tuunanen (2022). Responsible cognitive digital clones as decision-makers: A design science research study. *European Journal of Information Systems*. DOI: 10.1080/0960085X.2022.2073278. https://doi.org/10.1080/0960085X.2022.2073278

Marjanovic, O., Cecez-Kecmanovic, D., Vidgen, R. (2018). Algorithmic Pollution: Understanding and Responding to Negative Consequences of Algorithmic Decision-Making. In: Schultze, U., Aanestad, M., Mähring, M., Østerlund, C., Riemer, K. (eds) Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology. IS&O 2018. *IFIP Advances in Information and Communication Technology,* vol 543. Springer, Cham. https://doi.org/10.1007/978-3-030-04091-8_4

Martínez-Plumed F. et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," in *IEEE Transactions on Knowledge and Data*

Engineering, vol. 33, no. 8, pp. 3048-3061, 1 Aug. 2021, doi: 10.1109/TKDE.2019.2962680. https://doi.org/10.1109/TKDE.2019.2962680

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. arXiv. https://arxiv.org/abs/2412.04984

Merriam-Webster, Inc. (n.d.). *Algorithm*. In *Merriam-Webster.com dictionary*. Retrieved December 29, 2024, from https://www.merriam-webster.com/dictionary/algorithm

Minh, D., Wang, H.X., Li, Y.F. *et al.* Explainable artificial intelligence: a comprehensive review. *Artif Intell Rev* 55, 3503–3568 (2022). https://doi-org.proxy.uwasa.fi/10.1007/s10462-021-10088-y

Montanaro, U., Martini, S., Hao, Z., Gao, Y., & Sorniotti, A. (2023). Multi-input enhanced model reference adaptive control strategies and their application to space robotic manipulators. INTERN*ATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL,* 33(10), 5246-5272. https://doi.org/10.1002/rnc.6639

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M.-A. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. *Big Data Cogn. Comput.*, 5, 20. https://doi.org/10.3390/bdcc5020020.

Moore C., Ann E. Tenbrunsel (2014) “Just think about it”? Cognitive complexity and moral choice. *Organizational Behavior and Human Decision Processes*, Volume 123, Issue 2, 2014, Pages 138-149, ISSN 0749-5978, https://doi.org/10.1016/j.obhdp.2013.10.006.

Mufson, B. (2017, March 18). Meet the artist using ritual magic to trap self-driving cars. VICE. https://www.vice.com/en/article/qkmeyd/meet-the-artist-using-ritual-magic-to-trap-self-driving-cars

Musch, S., Borrelli, M., & Kerrigan, C. (2023). The EU AI Act as global artificial intelligence regulation. SSRN. https://ssrn.com/abstract=4549261

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance (arXiv:2206.00335). arXiv. https://doi.org/10.48550/arXiv.2206.00335

Naidu, G., Zuva, T., Sibanda, E.M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy, R., Silhavy, P. (eds) Artificial Intelligence Application in Networks and Systems. CSOC 2023. Lecture Notes in *Networks and Systems,* vol 724. Springer, Cham. https://doi.org/10.1007/978-3-031-35314-7_2

Nassereddine, M., Ellakkis, M.A., Azar, A. *et al.* Developing a Multi-methodology for Conflict Resolution: Case of Yemen's Humanitarian Crisis. *Group Decis Negot* **30**, 301–320 (2021). https://doi.org/10.1007/s10726-020-09695-x

Negar Maleki, Balaji Padmanabdhan and Kaushik Dutta, 2024: AI Hallucinations: A Misnomer Worth Clarifying. https://doi.org/10.48550/arXiv.2401.06796

Nevanperä Minna, Rajamäki Jyri & Helin Jaakko (2021) Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers. *25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*. 192 (2021) 2330–2339 http://dx.doi.org/10.1016/j.procs.2021.08.223

Nguyen, T. T., Hui, P. M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014, February). Exploring the filter bubble: the effect of using recommender systems on content diversity. *In Proceedings of the 23rd international conference on World Wide Web* (pp. 677-686). http://dx.doi.org/10.1145/2566486.2568012

Nguyen, H., Ahn, J., Belgrave, A., Lee, J., Cawelti, L., Kim, H. E., Prado, Y., Santagata, R., & Villavicencio, A. (2021). Establishing Trustworthiness Through Algorithmic Approaches to Qualitative Research. *Advances in Quantitative Ethnography*, 1312, 47-61. https://doi.org/10.1007/978-3-030-67788-6_4

Nripendra P. Rana, Sheshadri Chatterjee, Yogesh K. Dwivedi & Shahriar Akter (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems,* 31(3), 364-387. DOI: 10.1080/0960085X.2021.1955628. https://doi.org/10.1080/0960085X.2021.1955628

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining Knowl Discov.,* 10, e1356. https://doi.org/10.1002/widm.1356.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Olivera Marjanovic, Dubravka Cecez-Kecmanovic & Richard Vidgen (2022). Theorising Algorithmic Justice. *European Journal of Information Systems*, 31(3), 269-287. DOI: 10.1080/0960085X.2021.1934130. https://doi.org/10.1080/0960085X.2021.1934130

Olivera Marjanovic, Dubravka Cecez-Kecmanovic (2017). Exploring the tension between transparency and datification effects of open government IS through the lens of Complex Adaptive Systems. *The Journal of Strategic Information Systems*, 26(3), 210-232. https://doi.org/10.1016/j.jsis.2017.07.001.

Oswald M, Jamie Grace, Sheena Urwin & Geoffrey C. Barnes (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455 https://doi.org/10.1080/13600834.2018.1458455

Oxford University Press. (n.d.). *Algorithm*. In *Oxford English Dictionary*. Retrieved December 29, 2024, from https://www.oed.com/dictionary/algorithm_n?tl=true

Page, Xinru & Marabelli, Marco & Tarafdar, Monideepa. (2017). Perceived Role Relationships in Human-Algorithm Interactions: The Context of Uber Drivers. *ICIS 2017 Proceedings.* 25. https://aisel.aisnet.org/icis2017/HumanBehavior/Presentations/25

Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. In Proceedings of the *ACM Conference on Fairness, Accountability, and Transparency* (FAT* '19), January 29-31, 2019. ACM, Atlanta. https://doi.org/10.1145/3287560.3287567.

Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström & Aleš Popovič (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal*

of Information Systems, 31(3), 257-268. DOI: 10.1080/0960085X.2022.2026621.https://doi.org/10.1080/0960085X.2022.2026621

Pavlidis, G. (2024). Unlocking the black box: Analyzing the EU artificial intelligence act's framework for explainability in AI. *Law, Innovation and Technology*, 16(1), 293-308. https://doi.org/10.1080/17579961.2024.2313795

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Abi Raad, M., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., ... Shevlane, T. (2024). Evaluating frontier models for dangerous capabilities. arXiv preprint arXiv:2403.13793. https://doi.org/10.48550/arXiv.2403.13793

Pizarro-Vasquez, G. O., Morales, F. M., Minervini, P. G., & Botto-Tobar, M. (2020). Sorting Algorithms and Their Execution Times: An Empirical Evaluation. *Advances in Intelligent Systems and Computing,* 1302, 335-348. SpringerLink. https://doi.org/10.1007/978-3-030-63665-4_27

Peffers Ken, Tuure Tuunanen, Marcus Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. J. Manage. *Inf. Syst.* 24, 3 (Number 3 / Winter 2007-2008), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility. *Social Media + Society*, *5*(4). https://doi.org/10.1177/2056305119879995

Peruzzi A, Zollo F, Quattrociocchi W, Scala A. How News May Affect Markets' Complex Structure: The Case of Cambridge Analytica. *Entropy* (Basel). 2018 Oct 6;20(10):765. doi: 10.3390/e20100765. PMID: 33265853; PMCID: PMC7512327. https://doi.org/10.3390/e20100765

Prichard Eric C., 2021, Is the Use of Personality Based Psychometrics by Cambridge Analytical Psychological Science's "Nuclear Bomb" Moment?, *Frontiers in Psychology,* 12 https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.581448, DOI 10.3389/fpsyg.2021.58144

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

Pumplun, Luisa; Peters, Felix; Gawlitza, Joshua F.; and Buxmann, Peter (2023) "Bringing Machine Learning Systems into Clinical Practice: A Design Science Approach to Explainable Machine Learning-Based Clinical Decision Support Systems," *Journal of the Association for Information Systems,* 24(4), 953-979. DOI: 10.17705/1jais.00820 https://aisel.aisnet.org/jais/vol24/iss4/8

Pär J Ågerfalk, Kieran Conboy & Michael D Myers (2020). Information systems in the age of pandemics: COVID-19 and beyond. *European Journal of Information Systems*, 29(3), 203-207. DOI: 10.1080/0960085X.2020.1771968. https://doi.org/10.1080/0960085X.2020.1771968

Raden, N. (2020). Ethical Issues in any Automated Decision-Making Model. *Actuarial Technology Today*. Retrieved from https://www.soa.org/globalassets/assets/library/newsletters/actuarial-technology-today/2020/july/att-2020-07-raden.pdf

Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. Sci Rep 14, 6086 (2024). https://doi.org/10.1038/s41598-024-56706-x

Rajibul Hasan, Riad Shams, Mizan Rahman, (2021). Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research,* 131, 591-597. https://doi.org/10.1016/j.jbusres.2020.12.012.

Rambachan, Ashesh and Kleinberg, Jon and Mullainathan, Sendhil and Ludwig, Jens, 2020. An Economic Approach to Regulating Algorithms, *National Bureau of Economic Research,* 27111, doi:10.3386/w27111 https://ssrn.com/abstract=3597843

Ramón J.S., Daniel Palacios-Marqués, Agustín Iturricha-Fernández, Ethical design in social media: Assessing the main performance measurements of user online behavior modification, *Journal of Business Research*, Volume 129, 2021, Pages 271-281, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2021.03.001.

Reijonen Joel Master Thesis Decentralized Machine Learning for Autonomous Ships in Distributed Cloud Environment 2018

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior & Emerging Technologies, 1*. https://doi.org/[https://doi.org/10.1002/hbe2.117

Rinta-Kahila Tapani, Ida Someh, Nicole Gillespie, Marta Indulska & Shirley Gregor (2022) Algorithmic decision-making and system destructiveness: A case of automatic debt recovery, *European Journal of Information Systems*, 31:3, 313-338, DOI: 10.1080/0960085X.2021.1960905 https://doi.org/10.1080/0960085X.2021.1960905

Rubim Borges Fortes, P. (2020). Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process. *Asian Journal of Law and Society*, *7*(3), 453–469. doi:10.1017/als.2020.12 https://doi.org/10.1017/als.2020.12

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1, 206-215. arXiv:1811.10154 [stat.ML]. https://doi.org/10.48550/arXiv.1811.10154

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4), 105-114. https://doi.org/10.1609/aimag.v36i4.2577

Sá, Gabriel & Madeira, Charles. (2024). Deep reinforcement learning in real-time strategy games: a systematic literature review. *Applied Intelligence*. 55. 10.1007/s10489-024-06220-4. https://doi.org/10.1007/s10489-024-06220-4

Saltz JS (2021) Crisp-DM for data science: Strengths, weaknesses and potential next steps. In 2021 *IEEE International Conference on Big Data (Big Data).* Orlando, FL: IEEE, pp. 2337–2344 DOI: 10.1109/BigData52589.2021.9671634

Saltz JS, Krasteva I. (2022) Current approaches for executing big data science projects-a systematic literature review. *PeerJ Comput Sci.* 2022 Feb 21;8:e862. doi: 10.7717/peerj-cs.862. PMID: 35494858; PMCID: PMC9044260. https://doi.org/10.7717/peerj-cs.862

Saltz JS, and Neil Dewar. 2019. Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics and Inf. Technol*. 21, 3 (Sep 2019), 197–208. https://doi.org/10.1007/s10676-019-09502-5

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014, June). Auditing algorithms: Research methods for detecting discrimination on internet platforms. In Data and discrimination: converting critical concerns into productive inquiry (pp. 1-23). https://api.semanticscholar.org/CorpusID:15686114

Sanur Sharma (2023) Trustworthy Artificial Intelligence: Design of AI Governance Framework, *Strategic Analysis*, 47:5, 443-464, DOI: 10.1080/09700161.2023.2288994 https://doi.org/10.1080/09700161.2023.2288994

Samir Passi and Solon Barocas. (2019) Problem Formulation and Fairness. *In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAT* '19), January 29-31, 2019. ACM, Atlanta. https://doi.org/10.1145/3287560.3287567

Samuli Laato, A. K. M. Najmul Islam, Muhammad Nazrul Islam & Eoin Whelan (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, 29(3), 288-305. DOI: 10.1080/0960085X.2020.1770632. https://doi.org/10.1080/0960085X.2020.1770632

Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI*. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

Shaza H. Mansour, Sarah M. Azzam, Hany M. Hasanien, Marcos Tostado-Véliz, Abdulaziz Alkuhayli, Francisco Jurado, 2025, Deep reinforcement learning-based plug-in electric vehicle charging/discharging scheduling in a home energy management system, *Energy*, Volume 316, 2025, 134420,ISSN 0360-5442, https://doi.org/10.1016/j.energy.2025.134420.

Schelble, B. G., Flathmann, C., McNeese, N., & Canonico, L. B. (2021). Understanding human-AI cooperation through game-theory and reinforcement learning models. In *Proceedings of the Annual Hawaii International Conference on System*

Sciences (Vol. 2020-January, pp. 348–357). IEEE Computer Society. https://doi.org/10.24251/hicss.2021.041

Schuetz, S., & Venkatesh, V. (2020). Research Perspectives: The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *Journal of the Association for Information Systems*, 21(2). DOI: 10.17705/1jais.00608. https://ssrn.com/abstract=3680306

Schwartz, O. (2019, November 25). In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE Spectrum*. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Shah Miah, Jahan; Gammack, John G; McKay, Judy. (2019) A Metadesign Theory for Tailorable Decision Support. *Journal of the Association for Information Systems*; Atlanta Vol. 20, Iss. 5, (2019): 570-603. DOI:10.17705/1jais.00544 https://doi.org/10.17705/1jais.00544

Shao R, Shi Z, Zhang D. Social Media and Emotional Burnout Regulation During the COVID-19 Pandemic: Multilevel Approach. *J Med Internet Res*. 2021 Mar 16;23(3):e27015. doi: 10.2196/27015. PMID: 33661753; PMCID: PMC7968478. https://doi.org/10.2196/27015

Sharjeel Tahir, Syed Afaq Shah, Jumana Abu-Khalaf, (2023) Artificial Empathy Classification: A Survey of Deep Learning Techniques, Datasets, and Evaluation Scales. *CoRR* abs/2310.00010 (2023) DOI: 10.48550/ARXIV.2310.00010 https://doi.org/10.48550/arXiv.2310.00010

Sherry Y. and N. C. Thompson, "How Fast Do Algorithms Improve? [Point of View]," *in Proceedings of the IEEE,* vol. 109, no. 11, pp. 1768-1777, Nov. 2021, doi: 10.1109/JPROC.2021.3107219. https://doi.org/10.1109/JPROC.2021.3107219

Sholokhova, S., Bizzari, V. & Fuchs, T. Exploring phenomenological interviews: questions, lessons learned and perspectives. *Phenom Cogn Sci* 21, 1–7 (2022). https://doi.org/10.1007/s11097-021-09799-y

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI. *AMC Transactions on Interactive Intelligent Systems*, 26. DOI: https://doi.org/10.1145/3419764.

Shukla, S., Bisht, K., Tiwari, K., Bashir, S. (2023). Navigating the Data Deluge: Challenges and Opportunities. In: Data Economy in the Digital Age. *Data-Intensive Research*. Springer, Singapore. https://doi.org/10.1007/978-981-99-7677-5_2

Silvola Hanna, Landau Tiina, (2019), Vastuullisuudesta ylituottoa sijoituksiin. 1st edition. Alma Media pages 18-19

Sedgewick Robert & Wayne Kevin. (2011) Algorithms 4th edition. Addison Wesley pages 4–5.

Sein, Maung K.; Henfridsson, Ola; Purao, Sandeep; Rossi, Matti; and Lindgren, Rikard. 2011. "Action Design Research," *MIS Quarterly*, (35: 1) pp.37-56. http://dx.doi.org/10.2307/23043488

Sevilla J., Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos (20) Compute Trends Across Three Eras of Machine Learning. 2022 *International Joint Conference on Neural Networks (IJCNN),* Padua, Italy, 2022, pp. 1-8 Related DOI: https://doi.org/10.1109/IJCNN55064.2022.9891914

Smith CJ. Type I and Type II errors: what are they and why do they matter? *Phlebology*. 2012;27(4):199-200. doi:10.1258/phleb.2012.012j04 https://doi.org/10.1258/phleb.2012.012j04

Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. (2019) Model Reconstruction from Model Explanations. In FAT* ’19: *Conference on Fairness, Accountability, and Transparency* (FAT* ’19), January 29–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA, DOI: 10.1145/3287560.3287562 https://doi.org/10.1145/3287560.3287562

Stanford University Institute for Human Centered Artificial Intelligence (2025) retrieved 20.05.2024 available at https://hai.stanford.edu/navigate/welcome

Stelmaszak, Marta & Möhlmann, Mareike & Sørensen, Carsten. (2024). When Algorithms Delegate to Humans: Exploring Human-Algorithm Interaction at Uber. *MIS Quarterly*. 10.25300/MISQ/2024/17911. http://dx.doi.org/10.25300/MISQ/2024/17911

Su, J., Huang, J., Adams, S., Chang, Q., & Beling, P. A. (2022). Deep multi-agent reinforcement learning for multi-level preventive maintenance in manufacturing systems. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2021.116323.

Sue Newell, Marco Marabelli (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems*, 24(1), 3-14. https://doi.org/10.1016/j.jsis.2015.02.001.

Sun Heshan, Yulin Fang, Haiyun (Melody) Zou (2016) Choosing a Fit Technology: Understanding Mindfulness in Technology Adoption and Continuance. *Journal of Association for Information Systems.* Volume 17, issue 6, pp. 377-412 June 2016. ISSN: 1536-9323 http://dx.doi.org/10.17705/1jais.00431

Susan A. Brown, Viswanath Venkatesh, Sandeep Goyal (2012). Expectation Confirmation in Technology Use. *Information Systems Research*, 23(2), 474–487. http://dx.doi.org/10.1287/isre.1110.0357.

Spiegelhalter D (2020) Should We Trust Algorithms? *Harvard Data Science Review*, Issue 2, winter 2020. Published on: Jan 31, 2020. DOI: 10.1162/99608f92.cb91a35a https://doi.org/10.1162/99608f92.cb91a35a

Szymański, P., Rademakers, F., & Fraser, A. G. (2024). The Artificial Intelligence Act approved by the EU: The difficult dialogue between the black box and the cardiologist. *European Heart Journal*, 45(30), 2686-2688. https://doi.org/10.1093/eurheartj/ehae281

S.2770 - 118th Congress (2023-2024): Protect Elections from Deceptive AI Act. (2024, May 15). https://www.congress.gov/bill/118th-congress/senate-bill/2770

Tamò-Larrieux, A., Guitton, C., Mayer, S., & Lutz, C. (2023). Regulating for trust: Can law establish trust in artificial intelligence? *Regulation & Governance*, 18(3), 780-801. https://doi.org/10.1111/rego.12568

Tams, S., Thatcher, J. B., & Grover, V. (2018). Concentration, Competence, Confidence, and Capture: An Experimental Study of Age, Interruption-based Technostress, and Task Performance. *Journal of the Association for Information Systems*, 19(9), 857-908. doi: 10.17705/1jais.00511. http://dx.doi.org/10.17705/1jais.00511

Tae Rang Choi, Minette E. Drumwright (2021) “OK, Google, why do I use you?” Motivations, post-consumption evaluations, and perceptions of voice AI assistants. *Telematics and Informatics*, Volume 62, 2021, 101628, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2021.101628.

Tarafdar, M., Page, X., & Marabelli, M. (2023). Algorithms as co-workers: Human algorithm role interactions in algorithmic work. *Information Systems Journal*, 33(2), 232–267. https://doi.org/10.1111/isj.12389

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Q.,* 45. DOI:10.25300/misq/2021/16535. http://dx.doi.org/10.25300/MISQ/2021/16535

Tindall Joseph, Matthew Fishman, E. Miles Stoudenmire, Dries Sels. Efficient Tensor Network Simulation of IBM’s Eagle Kicked Ising Experiment. *PRX Quantum*, 2024; 5 (1) DOI: 10.1103/PRXQuantum.5.010308 https://doi.org/10.1103/PRXQuantum.5.010308

Tong, S., Luo, X. & Xu, B. Personalized mobile marketing strategies. *J. of the Acad. Mark. Sci.* 48, 64–78 (2020). https://doi.org/10.1007/s11747-019-00693-3

Tsamados, A., Aggarwal, N., Cowls, J. *et al.* The ethics of algorithms: key problems and solutions. *AI & Soc* **37**, 215–230 (2022). https://doi.org/10.1007/s00146-021-01154-8

Tsiakas, Konstantinos & Murray-Rust, Dave. (2024). Unpacking Human-AI interactions: From Interaction Primitives to a Design Space. *ACM Transactions on Interactive Intelligent Systems*. 14. 10.1145/3664522. https://doi.org/10.48550/arXiv.2401.05115

UK Government. (2023). *The Bletchley Declaration by countries attending the AI Safety Summit (1–2 November 2023)*. Retrieved from https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

US Department of Justice, Hackers Use Deep Voice Tech in $35 Million Theft, Case 1:21-ml-00887-RBC (2021). https://www.documentcloud.org/documents/21085009-hackers-use-deep-voice-tech-in-400k-theft DOJ Ref. # CRM-182-77215

Xing, B., & Marwala, T. (2017). Introduction to Intelligent Search Algorithms. *Studies in Systems, Decision and Control,* 129, 33-64. SpringerLink. https://doi.org/10.1007/978-3-319-67480-3_3

Xu, Wei & Dainoff, Marvin & Ge, Liezhong & Gao, Zaifeng. (2021). From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI. https://doi.org/10.48550/arXiv.2105.05424

Yanqing Duan, John S. Edwards & Yogesh K Dwivedi (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management* 48 (2019) 63–71 https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Yao, S., Yu, D., Shi, X., Zhao, Y., Cao, Y., Narasimhan, K., & Liang, P. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint arXiv:2305.10601. https://doi.org/10.48550/arXiv.2305.10601

van de Poel, I. (2023). AI, Control and Unintended Consequences: The Need for Meta-Values. In: Fritzsche, A., Santa-María, A. (eds) Rethinking Technology and Engineering. *Philosophy of Engineering and Technology*, vol 45. Springer, Cham. https://doi.org/10.1007/978-3-031-25233-4_9

Wang, J., Li, Y., & Rao, H. R. (2016). Overconfidence in Phishing Email Detection. *Journal of the Association for Information Systems,* 17(11). DOI: 10.17705/1jais.00442. http://dx.doi.org/10.17705/1jais.00442

Wang, Y., He, H., & Sun, C. (2018). Learning to Navigate Through Complex Dynamic Environment With Modular Deep Reinforcement Learning. *IEEE Transactions on Games,* 10(4), 400-412. doi: 10.1109/TG.2018.2849942. https://doi.org/10.1109/TG.2018.2849942

Wang, Di & Zheng, Kaiyang & Li, Chuanni & Guo, Jianting. (2024). Transitioning to Human-Centered AI : A Systematic Review of Theories, Scenarios, and Hypotheses in Human- *AI Interactions. Proceedings of the Association for Information Science*

and Technology. 61. 673-678. 10.1002/pra2.1078. http://dx.doi.org/10.1002/pra2.1078

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint arXiv:2203.11171. https://doi.org/10.48550/arXiv.2203.11171

Warkentin, M., Walden, E., Johnston, A. C., & Straub, D. W. (2016). Neural Correlates of Protection Motivation for Secure IT Behaviors: An fMRI Examination. *Journal of the Association for Information Systems*, 17(3). DOI: 10.17705/1jais.00424. http://dx.doi.org/10.17705/1jais.00424

Watson, H. J. (2017). Preparing for the Cognitive Generation of Decision Support. *MIS Q. Executive*, 16. https://www.researchgate.net/publication/319929429.

Watson, H. J., & Nations, C. (2019). Addressing the Growing Need for Algorithmic Transparency. *Communications of the Association for Information Systems*, 45, pp-pp. https://doi.org/10.17705/1CAIS.04526.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? arXiv preprint arXiv:2307.02483. https://doi.org/10.48550/arXiv.2307.02483

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903. https://doi.org/10.48550/arXiv.2201.11903

Wei, K. K., Teo, H. H., Chan, H. C., & Tan, B. C. Y. (2011). Conceptualizing and Testing a Social Cognitive Model of the Digital Divide. *Information Systems Research*, 22(1), 170–187. doi: 10.1287/isre.1090.0273. http://dx.doi.org/10.1287/isre.1090.0273

Wilson, B.M., Harris, C.R. & Wixted, J.T. Theoretical false positive psychology. *Psychon Bull Rev* 29, 1751–1775 (2022). https://doi.org/10.3758/s13423-022-02098-w

Wills, P., & Meyer, F. G. (2020). Metrics for graph comparison: A practitioner's guide. *PLOS ONE*, 15(2), e0228728. https://doi.org/10.1371/journal.pone.0228728

White House. (2024, July 26). Fact sheet: Biden-Harris administration announces new AI actions and receives additional major voluntary commitment on AI [Press

release]. https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/26/fact-sheet-biden-harris-administration-announces-new-ai-actions-and-receives-additional-major-voluntary-commitment-on-ai/

Wu, Y., Choi, B., Guo, X., & Chang, K. T. (2017). Understanding User Adaptation toward a New IT System in Organizations: A Social Network Perspective. *Journal of the Association for Information Systems*, 18(11). DOI: 10.17705/1jais.00473. http://dx.doi.org/10.17705/1jais.00473

Wu W. et al. Wenjun Wu a,⇑, Tiejun Huang b, Ke Gong/ *Engineering* 6 (2020) 302–309 Ethical Principles and Governance Technology Development of AI in China https://doi.org/10.1016/j.eng.2019.12.015

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132. https://doi.org/10.1177/0162243915605575

Zhang, Shihao & Li, Lvzhou. (2022). A brief introduction to quantum algorithms. 10.48550/arXiv.2212.10734. https://doi.org/10.1007/s42514-022-00090-3

Zhang, Y., & Zhao, Q. (2024). Complex Environment Based on Improved A* Algorithm Research on Path Planning of Inspection Robots. *Processes*, 12(5), 855. https://doi.org/10.3390/pr12050855.

Zheng, J. (Frank), & Jarvenpaa, S. (2021). Thinking Technology as Human: Affordances, Technology Features, and Egocentric Biases in Technology Anthropomorphism. *Journal of the Association for Information Systems*, 22(5), 1429-1453. DOI: 10.17705/1jais.00698. Available at: https://aisel.aisnet.org/jais/vol22/iss5/3.

Ziewitz, M. (2019). Rethinking gaming: The ethical work of optimization in web search engines. Social Studies of Science, 49(5), 707-731. https://doi.org/10.1177/0306312719865607