NPL
National Physical Laboratory

NPL Report MS 57

A LIFE CYCLE FOR TRUSTWORTHY AND SAFE ARTIFICIAL INTELLIGENCE SYSTEMS

MARK LEVENE, TAMEEM ADEL, MOULHAM ALSULEMAN, INDHU GEORGE, PADMINI KRISHNADAS, KEITH LINES, YUHUI LUO, IAN SMITH AND PAUL DUNCAN

APRIL 2024

Department for
Science, Innovation
& Technology

A Life Cycle for Trustworthy and Safe Artificial Intelligence Systems

Mark Levene, Tameem Adel, Moulham Alsuleman, Indhu George, Padmini Krishnadas, Keith Lines, Yuhui Luo, Ian Smith and Paul Duncan*

Data Science Department

*paul.duncan@npl.co.uk

ABSTRACT

This work presents a life cycle for trustworthy and safe artificial intelligence systems (TAIS). We consider the application of a risk management framework (RMF) in the TAI life cycle, and how this affects the choices made during the various phases of AI system development. The emerging requirements for AI systems go beyond the traditional software engineering (SE) development cycle of design, develop and deploy (DDD), where test, evaluation, verification and validation (TEVV) plays a crucial role. In particular, the additional challenge in SE is that AI systems are, in general: (i) data-driven, (ii) can learn and (iii) are adaptive. Moreover, considerations of the impact of AI systems on users and wider society are essential to their trustworthy and safe deployment. In order to achieve a TAIS we also consider the measurement aspect, which manifests itself in the specification of quantifiable metrics capturing both technical and socio-technical principles of TAI that allow us to assess the trustworthiness of an AI system. The TAI life cycle is an iterative and continuous process, so that the AI system being developed is evolving and corrective action can be taken to deal with arising issues at any stage of the life cycle.

NPL Report MS 57

ISSN 1754-2960

https://doi.org/10.47120/npl.MS57

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

This work was funded by the UK Government's Department for Science, Innovation & Technology through the UK's National Measurement System programmes.

Approved on behalf of NPLML by Peter Harris,

Science Area Leader, Data Anayltics & Modelling, Data Science

ACRONYMS

AI Artificial Intelligence.

CDEI Centre for Data Ethics and Innovation; called Responsible Technology Adoption Unit (RTA) from February 2024.

CT Computed Tomography.

DDD Design, Develop and Deploy.

DNN Deep Neural Network.

E2E End-to-End.

GAI Generative AI.

GDPR General Protection Data Regulation.

HITL Human-In-The-Loop.

ISO International Organization for Standardization.

LIME Local Interpretable Model-agnostic Explanations.

LLM Large Language Model.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

NIST National Institute of Standards and Technology.

NPL National Physical Laboratory.

OECD Organisation for Economic Co-operation and Development.

R3 Reliability, Resilience and Robustness.

RMF Risk Management Framework.

SE Software Engineering.

SWEBOK Software Engineering Body of Knowledge.

TAI Trustworthy and safe Artificial Intelligence.

TAIS Trustworthy and safe Artificial Intelligence System.

TEVV Test, Evaluation, Verification and Validation.

UK United Kingdom of Great Britain and Northern Ireland.

UQ Uncertainty Quantification.

XAI eXplainable and interpretable AI.

CONTENTS

TABLES

FIGURES

1 EXECUTIVE SUMMARY

Artificial intelligence (AI) is recognised as one of the key transformative technologies promising to deliver strong positive impact in key areas, such as healthcare, autonomous vehicles, energy and climate, where the National Physical Laboratory (NPL) is providing cutting-edge measurement science to underpin the prosperity and quality of life in the UK. AI systems also pose risks, some of which may be serious in terms of the harm that may be caused to individuals or society [41, 44], and therefore to realise their promise AI systems need to be trustworthy in the sense that they can be relied upon to make responsible and safe decisions [37].

It is important to understand that AI is a broader topic than machine learning (ML), since it includes a wider set of technologies that may not be considered ML as such. ML is a subset of AI concentrating on AI methods that are potentially able to learn and adapt, involving the construction of statistical models of data that may be used for classification and prediction tasks to aid decision-making. AI also includes symbolic computation, such as rule-based systems and knowledge graphs, which are not a part of ML. Moreover, an AI system may also involve a level of autonomous behaviour, a point which is explicitly mentioned in the OECD definition of AI [46]. Our focus here will be on AI systems which include the use of ML, but do not exclude the use of broader AI methods.

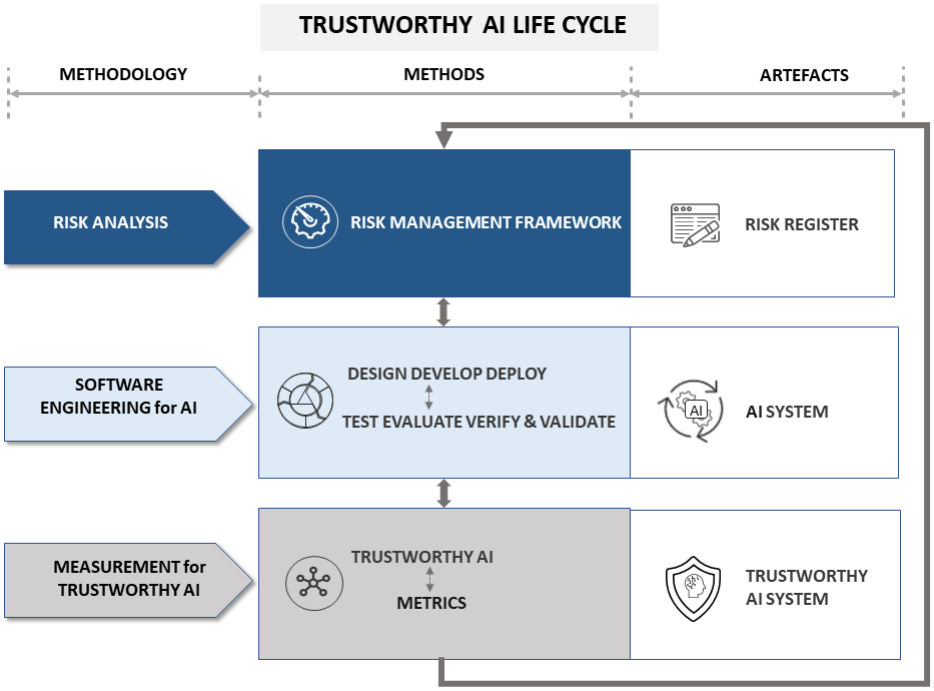

Figure 1: TAI life cycle diagram.

The objective in this work is to propose a life cycle for trustworthy and safe AI systems (TAIS), so that we can mitigate against the risks that an AI system may pose and maximise the benefits from its potential positive impacts. The suggested TAI life cycle diagram is

shown in Figure 1, observing that our primary focus is on software development, yet we do not rule out the inclusion of hardware components in the AI system. It is important to note that the arrows in the diagram indicate that the process underlying the life cycle is iterative and continuous allowing transitions between the three layers in either direction, bearing in mind that any change in one layer may have an effect on one or more of the other layers. Although we view each layer in the life cycle in terms of its *methodology*, the *methods* employed and the *artefacts* output, we may also use the more traditional labels for the columns as *stage* for the first, *processes* for the second and simply *outputs* for the third; see the ISO standard for AI system life cycle processes [29]. We emphasise that the life cycle describes the evolution of an AI system from inception to retirement [29]. The novelty of our proposal is that the TAI life cycle, shown in Figure 1, provides a high-level framework with three well-defined layers, such that when drilling down into each layer we are able to reuse or extend existing methods, and add new ones where necessary. Ultimately, our AI expertise in NPL is focused on measurement for TAI, where we are able to provide more innovation than in the other layers.

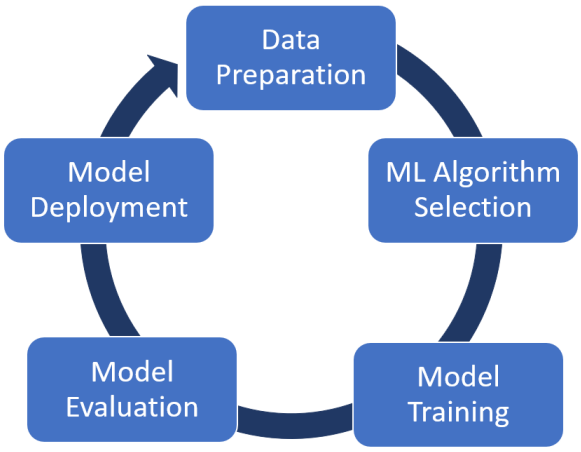

Figure 2: A simplified ML workflow, describing the sequence of steps in developing ML software.

Before we delve into the details of the diagram, we highlight the fact that AI systems differ from conventional deterministic computer systems in the sense that many are probabilistic in nature. Concentrating for the moment on the ML component of an AI system, we stress that it is different from conventional software in several aspects (see the simplified ML workflow diagram, shown in Figure 2; cf. [4]):

1) ML is data-driven. An ML algorithm constructs a statistical model from its input training data, and the accuracy of its outputs are evaluated from the test data. In addition, data quality is crucial during the data preparation stage.

2) ML can learn from input data. This manifests itself in the iterative and continuous nature of the workflow.

3) ML is potentially adaptive. In principle, an ML algorithm can adapt its model to take into account new data. This process may cause the model to change its response to inputs, and in this sense ML is inherently non-deterministic.

In the following we refer to the differences between AI and conventional systems as the *distinguishing features of AI Systems*. Moreover, we emphasise that AI systems may also

exhibit autonomous behaviour with a varying degree of human interaction, known as human-in-the-loop (HITL) [63], and it is often the case that autonomous behaviour is enabled by ML. Having HITL interaction may be critical to ensure the trustworthiness of the AI system, depending on the application; for example, consider the domain of AI in healthcare.

The first layer in Figure 1 is concerned with the activity of risk analysis for identifying risks and assessing their potential likelihood and consequences [26]. This activity is achieved with a risk management framework (RMF), which, upon recognising the risks and benefits arising from the AI system being developed, will recommend the necessary steps to minimise the negative impacts and, respectively, maximise the positive ones. We concentrate on risk management for AI systems, whose goal is to provide a pathway for designing, developing and deploying trustworthy and safe AI systems. In particular, we single out NIST's AI RMF [57] (which, for brevity, we often refer to simply as the RMF), although others have been proposed and will be compared to NIST's AI RMF in Section 4.

The core of the RMF comprises four functions: Govern, Map, Measure and Manage [57], as shown in Figure 3. Each function is broken down into subcategories, containing actions and outcomes to assist organisations to manage AI risks and undertake the responsible development and use of AI systems.

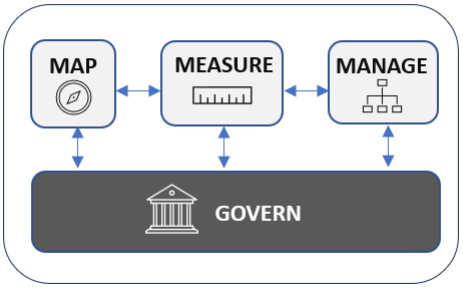

Figure 3: The core functions of NIST's AI RMF.

In particular, the Govern function deals with how all aspects of risk are managed within an organisation. The Map function establishes the context in which to frame the risks of an AI system. The Measure function uses information gathered from the Map function, together with specific tools and techniques, to monitor and analyse the risks. Finally, the Manage function allocates risk management resources to prioritise the risks identified by the Map and Measure functions on an ongoing basis.

The artefact output during the risk analysis stage, resulting from the risk management process, is the risk register [62], which acts as the project repository. The risk register catalogues the details of the identified risks and any other tangible outputs such as predictions, recommendations and decisions made during the risk management process.

The second layer in Figure 1 is concerned with the activity of software engineering (SE) for AI, which necessitates adjusting to account for the distinguishing features of AI systems. Here we will focus our attention on the design, develop and deploy (DDD) and test, evaluation, verification and validation (TEVV) cycles, which can be described as a coupled iterative process, as shown in Figure 4.

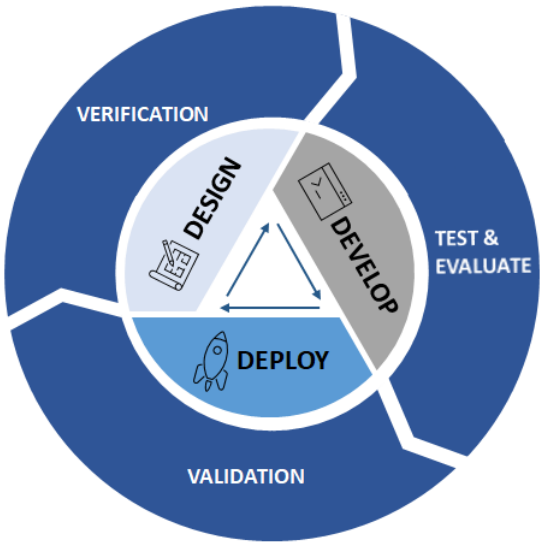

Figure 4: DDD-TEVV coupled cycles.

A comprehensive workflow for DDD in the context of AI is presented in [13], and can be viewed as a detailed version of the simplified ML workflow, shown in Figure 2. Regarding TEVV for AI, we first observe that validation is concerned with model evaluation, guaranteeing that the conditions required for deployment are satisfied, while verification is to do with model design. Furthermore, with regards to test and evaluation, there is a need to take into account the distinguishing features of AI systems [11, 25]. When we test a conventional algorithm against an input, then an error implies a "bug" in the code. For example, if the algorithm is meant to sort a set of numbers in ascending order, then there is only one correct result. On the other hand, in AI systems new data may change the result. In particular, the accuracy of an AI system is measured by the proportion of correct decisions rather than insisting that each decision is correct. It is also important to mention data quality issues [61], which can be associated with all aspects of TEVV and also with the evaluation of the interaction of humans with the AI system [5].

The artefact output during the SE stage, resulting from the DDD-TEVV process, is an AI system, which is to be constrained by TAI metrics at the third stage. It is important to note that the life cycle is iterative and continuous at all stages, so we are unlikely to deploy an AI system until it is considered to be sufficiently trustworthy and safe, resolving the catalogued risks.

The third layer in Figure 1 is concerned with the activity of measurement for TAI. The word trustworthy means "able to be relied upon as honest, responsible and truthful", which leads us to the concept of *trustworthy AI* [37]. We add safety to this list to emphasise the risks that AI poses and the importance of avoiding the potential harmful consequences, at various levels of severity, that may occur from the use of AI systems [41, 44]. Thus, by *TAI* we mean both trustworthy and safe AI.

The principles of TAI can be broken down into three pillars, as shown in Figure 5:

1) *Technical*, such as robustness, which is the technical system requirement of high performance under a wide range of conditions such as an unexpected or erroneous input; for example, in ML we often consider robustness to noise perturbations in the input.

2) *Socio-technical*, such as algorithmic fairness, which aims to ensure that the outputs of an AI system are fair, that is, they are free of bias and discrimination.

3) *Social*, such as accountability, which ensures that it is clear who is responsible for the outputs of an AI system.

We emphasise the socio-technical nature of AI systems, and that the boundary between the three pillars is, in many cases, unclear. This is highlighted in the double-arrowed line at the bottom of Figure 5, indicating that there is a large spectrum between technical and social. As a case in point fairness has a clear social meaning, while the socio-technical principle of algorithmic fairness is to ensure that the output of an AI system is fair, that is, free from bias and discrimination. (In the following we will generally refer to algorithmic fairness simply as fairness.) Detail on the principles shown in Figure 5 will be given in Section 6.

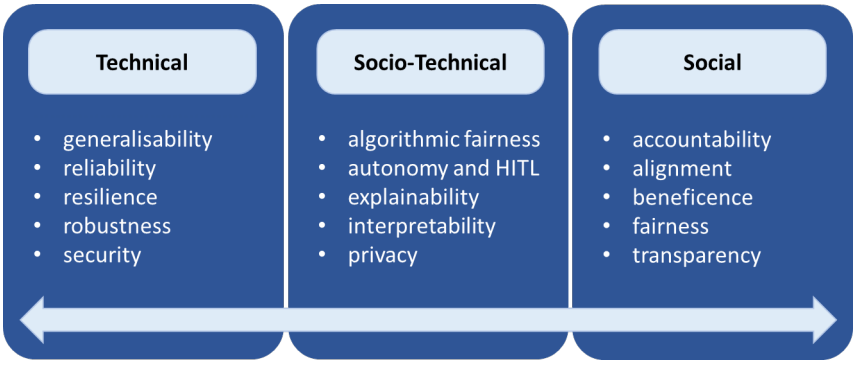

Figure 5: The three pillars of TAI principles.

Technical and socio-technical TAI principles may be operationalised with the aid of metrics. Thus by specifying quantifiable metrics for those principles that are relevant to the application at hand, we will be able to incorporate TAI into AI systems in a principled manner. Metrics will be discussed in Section 7, concentrating on interpretability, explainability, privacy and algorithmic fairness, which are important principles in many AI applications.

Finally, the artefact output during the measurement stage, resulting from the process that specifies the relevant TAI metrics, is a TAIS, which is an AI system constrained by the TAI metrics that pertain to the application at hand.

As the TAI life cycle is an iterative and continuous process, the artefacts produced may be refined at all stages of the software development process. This refinement is particularly true for AI systems, due to the fact that the constructed model may change over time as the system adapts to new data.

In a general sense, the audience of the TAI life cycle are organisations and individuals or teams developing AI systems. More specifically, the TAI life cycle is relevant not only to developers, scientists and researchers involved in the development of AI systems, but also to their managers at a higher level of the organisation, and to the end-users interacting with AI systems.

The rest of the white paper is organised as follows. In Section 2 we discuss the meaning of an end-to-end AI system, settling on an interpretation that is clear and appropriate in the context of this work. In Section 3 we briefly discuss three examples from the medical domain, to motivate the need for principled development of AI systems according to the TAI life cycle we propose. In Section 4 we concentrate on AI risk management, the first stage of the TAI life cycle. In Section 5 we concentrate on DDD-TEVV for AI systems, the second stage of the TAI life cycle. In Sections 6 and 7 we concentrate, respectively, on the pillars and principles of TAI and on the specification of selected metrics. In Section 8 we give our concluding remarks, and finally in Appendix A we present a taxonomy mapping of the principles and guidance recommended in four representative AI frameworks, which will help the reader understand the variations in terminology used for trustworthy and safe AI systems.

2 WHAT IS AN END-TO-END AI SYSTEM?

Much of the literature on AI systems describes those systems as being end-to-end, from the raw input data to its output, for example, [3, 56]. Given the widespread use of the adjective, which we abbreviate to E2E, we consider it relevant at this point to discuss briefly its meaning and the usefulness of the categorisation of AI systems as E2E.

Frequently, reference is made to an E2E AI system, for example, in the title or within the keywords of a paper, without the provision of a clear, explicit definition of the properties of such a system. A common interpretation involves employing a single AI component that takes source data as input and generates a final output. However, in another interpretation, sometimes qualified as modular, a series of separately trained AI components are used to go from source data to final output via intermediate outputs, for example, [38].

While these interpretations may be valid on their own, they are clearly contradictory. Indeed, an E2E AI system as given by the second interpretation could be considered to comprise a series of E2E AI systems as given by the first interpretation. Moreover, an AI system as a complete software system, will most likely include non-AI components.

There is also a lack of clarity about the role of humans in an E2E AI system. The first interpretation suggests no human involvement. For the second interpretation, there might be the risk of errors propagating through the system and therefore it could be considered useful to permit human intervention to prevent this occurrence. It therefore may be valid to amend the second interpretation and define an E2E AI system as one involving multiple components, at least one of which involves the use of AI.

The categorisation of an AI system as E2E is particularly interesting when considering medical applications where the aim is to diagnose conditions, for example, based on scanned images. In some cases, an AI system may be used to deliver a diagnostic opinion, but there is still the need for a human to use their expertise to confirm, or otherwise, that diagnosis [51]. There may be other factors, for example, a patient's medical records, that the AI system is not aware of and that can only be properly taken account of by a human.

The lack of precision regarding what an E2E AI system is, from the perspective of software development, suggests that use of the term should either be avoided or, if it is used, it should be accompanied by a clear statement of its interpretation in the context of its use. While no consensus definition of an E2E AI system as a whole exists, the term will not subsequently be used in this white paper. For simplicity, we will only use the term AI system to refer to a single AI component (that may or may not constitute the entire system and could conceivably include AI subcomponents) and we will continue using Figure 1 as a clearer representation of the iterative and continuous life cycle of AI systems.

3 EXAMPLES FROM THE MEDICAL DOMAIN

The healthcare domain is one of the strategic areas which NPL works on, aligning itself with the UK's strategic priorities. In particular, incorporating trustworthiness and patient safety into medical AI systems is of critical importance [52]; see the third layer of Figure 1. An interesting point here is that both the patients and the clinicians need to gain trust in the AI system. Amongst other applications, AI systems are employed in these settings to aid with diagnostic decision-making, treatment recommendations and personalised medicine.

For these applications, the use of AI systems in practice promises improved efficiency and accuracy, and enhanced diagnosis capabilities.

Translational medical AI is an important concept that is concerned with bridging the gap between research, which deals with the technical possibilities of the AI, and the design, development and deployment of real-world AI systems to support decision-making; see the second layer of Figure 1. An evaluation framework consisting of three main components, which focuses on the translational characteristics of AI systems, was proposed by [53]. They are: (i) capability of the AI system to perform as expected from a software (and hardware) development perspective, (ii) utility of the AI system, including usability, efficiency, safety and ethics considerations regarding the deployment of the system, and (iii) adoption and integration of the AI in a real-world healthcare setting.

In terms of risk management, medical AI amplifies certain risks, three of which are discussed in [6]; see the first layer of Figure 1. They are: (i) system malfunction, when the failure of an AI system can have a potentially devastating effect on patient care, (ii) privacy protection, which may be drastically more insecure due to AI, and, (iii) consent to data repurposing, which may be difficult to obtain but is necessary for AI's data-driven model.

We choose three examples, focusing on medical image analysis for radiology, to illustrate the need for principled development of AI systems according to the TAI life cycle we put forward.

The first example, [7], proposes a novel interpretable and explainable neural network algorithm that uses case-based reasoning [39] for mammography. The need for explainability and interpretability in AI (known as XAI; see Subsection 7.1) is addressed, when making high-stakes decisions in the medical domain such as the prescription of invasive surgeries. XAI outputs can be used by a radiologist to inform their decision-making.

The second example, [65], proposes a single pipeline deep learning framework for highly accelerated MRI image reconstruction and pathology detection. Insights into real-world challenges are provided, primarily for dealing with large quantities of raw data, and optimising the workflow, so that two or more reconstruction algorithms can be compared in terms of their detection capability.

Deep learning methods for MRI reconstruction were reviewed in [48] in a wide range of applications, especially in accelerated MRI acquisition and reconstruction tasks. The need for robust and accurate reconstruction of images with the relevant clinical details required for pathology detection is highlighted. There is also a need for the medical community to collaborate and collate both normative and pathological data sets for training the AI models.

The third example, [56], describes a deep learning method for automatic diagnosis of pancreatic tumours, from original abdominal CT images. The model provides an interpretable diagnosis for clinicians by highlighting the regions relevant to its decision.

All three examples demonstrate the viability of using AI software in the medical domain. However, in all three cases HITL interaction will need to be present to ensure the trustworthiness of the AI, despite the high accuracy of the initial results. We further remark that different domains, such as medical, legal and financial, will most likely have domain-specific requirements, which should be taken into consideration when developing a TAIS.

In the following sections we will dig deeper into the three layers of the TAI life cycle, bearing in mind that the development of an AI system to be deployed in the real world will involve an iterative and continuous enactment of the life cycle's layers.

4 UNDERSTANDING AI RISK MANAGEMENT

The understanding and management of risks is a key component of responsible development and use of AI systems; see [31]. These risks can range from inaccurate data leading to false results to systemic biases impacting each output of an AI system. Understanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust. Therefore, it is important that people who design, develop, and deploy AI systems think critically about the context and potential of risk to their systems.

As mentioned in Figure 1, the TAI life cycle contains three main elements: the methodology, the methods employed and the artefacts produced. We next discuss each of these elements in the context of the RMF [57]; see Figure 6.

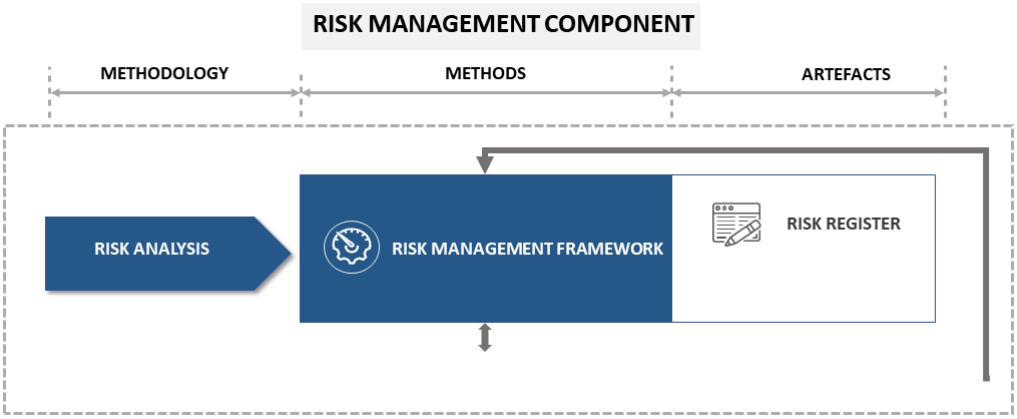

Figure 6: The RMF component of the TAI life cycle, shown in Figure 1.

4.1 METHODOLOGY

When assessing the risks of AI systems, the methodology we consider is *risk analysis* [26], i.e. how do we identify and evaluate which risks are associated with our system, with evaluation covering both how likely is the risk and what is its impact.

Risk assessment is performed throughout the AI system life cycle to ensure risk management is continuous and timely. Analysing and assessing risks in AI system development involves a multifaceted approach, requiring a thorough examination of technical, ethical, and societal implications. Technical aspects, such as security and data quality issues, must be considered alongside ethical concerns including privacy violations and discrimination. Societal risks, including job displacement and economic inequality, also demand attention. This process entails interdisciplinary collaboration, drawing expertise from computer science, ethics, law, and sociology. Risk management frameworks and methodologies tailored to AI development aid in identifying vulnerabilities and estimating their likelihood and impact. Proactive risk management strategies, such as comprehensive testing of the AI system and

continuous monitoring to identify new risks, are crucial to ensure responsible AI system development, promoting reliability, safety, and ethical integrity.

4.2 METHODS

To aid the assessment and understanding of risk in the context of AI system development, there have been several national and international developments in this space including: CDEI's AI assurance roadmap [10], the European AI Act [16, 44], ISO reports and standards [31, 32], NIST's AI RMF [57], as well as OECD's Tools for trustworthy AI [45], all shown in Table 1.

Table 1: High-level comparison of different AI frameworks.

FRAMEWORK	SCOPE	BENEFITS for TAISs
CDEI AI assurance roadmap [10]	UK/Voluntary	AI assurance
EU AI Act [16, 44]	EU/Mandatory	Regulatory compliance
ISO/IEC 23894 [31] and 42001 [32]	Global/Voluntary	AI risk and system management
NIST AI RMF [57]	Global/Voluntary	AI risk management
OECD Tools for trustworthy AI [45]	Global/Voluntary	Framework for implementation of tools

Each of these frameworks takes a particular approach to risk management of AI systems for specific applications. In Appendix A we briefly discuss how the four organisations, i.e. the CDEI, EU, NIST and OECD, treat the key high-level terms relevant to TAI.

For understanding the risk element of AI system development we concentrate on NIST's AI Risk Management Framework [57], which we will refer to here as the RMF. This is due to its general, detailed and flexible nature, which we will now elaborate on.

The RMF, released by the National Institute of Standards and Technology (NIST) in January 2023 [57], offers a comprehensive resource to organisations designing, developing, deploying, or using AI systems to help manage the risks of AI and promote responsible development and use of AI systems. The RMF is intended to be voluntary, rights-preserving, non-sector-specific, and use-case agnostic, providing flexibility to organisations of all sizes and in all sectors and throughout society to implement the approaches in the RMF.

It is important to note that the RMF is also not prescriptive on what tests should be carried out to monitor the system risks, and what controls are needed to maintain or modify identified risks. That is, the user of the RMF will still need to decide how it will be operationalised in practice.

Following this process for managing AI risks has the intention of maximising the benefits of AI technologies while reducing the likelihood of negative impacts to individuals, groups, communities, organisations and society. The RMF is designed to prepare AI actors (i.e. organisations and individuals) with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment and use of AI systems over time.

To understand the role of risk and risk management in the AI life cycle presented in this work, we first must examine how risk is assessed using the *method* of the RMF. Figure

3 shows the four main functions within the RMF. These key functions, namely, Govern, Map, Measure, and Manage are broken down further into categories and subcategories. While the Govern function applies to all stages of an organisation's AI risk management processes and procedures, the Map, Measure, and Manage functions can be applied in AI system-specific contexts and at specific stages of the AI life cycle.

Govern: The foundational function of the RMF. It builds a culture of risk management, defines processes, and provides a structure to the program by developing policies, processes and procedures to enable a structure around AI accountability.

Map: Establishes AI system context and context related risks. The Map function identifies the capability of the AI system and associated impacts of risks to individuals, groups, communities and organisations.

Measure: Analyses, assesses, benchmarks, and monitors the identified AI risks and related impacts by evaluating AI systems for trustworthy characteristics and creating mechanisms to track identified AI risks over time. A final part of the Measure function is to gather feedback about the efficacy of these measurements over time, ensuring that as the AI system evolves, so does the associated description of risk.

Manage: Prioritises and acts upon risks based on their projected impact, by identifying strategies to maximise AI benefits and minimise negative impacts. The Manage function also manages risks from third-party AI systems and identifies risk treatments, including response and recovery plans. For an organisation, assuming the governance structure is set up and in place, the other functions may be performed in any order across the AI life cycle, as such a strategy is considered to add value to the users of the RMF. In whatever way users integrate the functions of the RMF, the process should be iterative and continuous, with cross-referencing between functions as required.

4.3 Artefacts

A risk register (or project risk register) is a document which is kept under strict configuration control [62]. The *risk register* artefact is a list/log/repository of adverse events that might occur in the AI system. According to NIST, the adverse events are the potential harms to people, organisations or systems resulting from developing and deploying AI systems [57]. The risk register will also document the steps, decisions and actions taken during risk management. Using a risk register for AI system development is essential for systematically identifying, documenting, and managing potential hazards throughout the project life cycle. The risk register serves as a centralised repository that categorises risks based on their nature, likelihood, and potential impact. It allows project teams to prioritise mitigation efforts effectively by allocating resources to address the most critical risks. Regular updates to the risk register enable tracking of risk status, monitoring of mitigation measures, and communication of risk information across stakeholders. By incorporating feedback from various stakeholders and leveraging historical data, the risk register facilitates informed decision-making and proactive risk management, ultimately enhancing the reliability, safety, and ethical integrity of AI systems.

Risk management for AI systems is a vital element of the TAI life cycle as it allows users to identify, mitigate and track all risks pertaining to the system development. As in Figure 1 we are considering this process to be iterative and continuous due to the evolving nature of AI systems, meaning that the risks will continue to evolve alongside the system in deployment.

5 UNDERSTANDING DDD-TEVV FOR AI

We now discuss the three main elements of the TAI life cycle, as shown in Figure 1, that is, the methodology, the methods employed and the artefacts produced, in the context of the DDD-TEVV; see Figure 7. We observe that TEVV processes are performed throughout the TAI life cycle as part of its iterative and continuous nature, and that they assist in the risk management process.

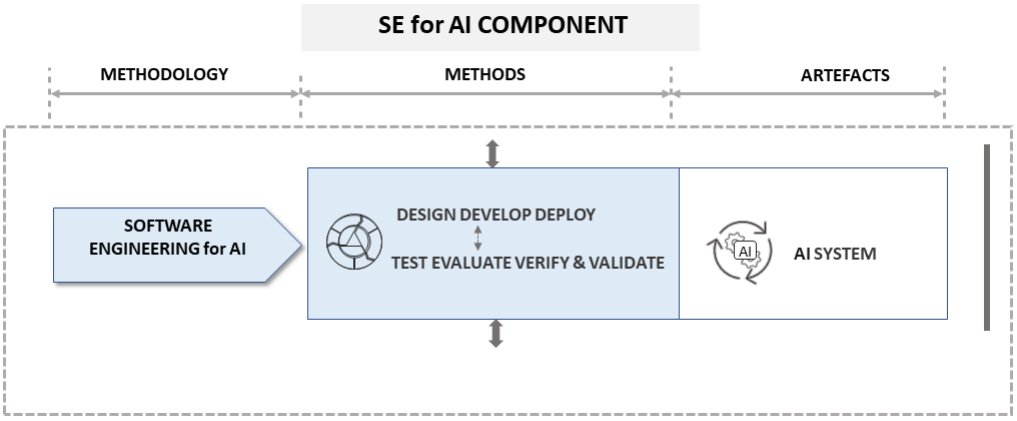

Figure 7: The DDD-TEVV component of the TAI life cycle, shown in Figure 1.

5.1 METHODOLOGY

SE for designing, developing and deploying AI systems needs to take into account the distinguishing features of AI systems. However, we first need to ascertain the potential usefulness of SE as a discipline for developing AI software. To answer this question [42] formulated five laws.

The first law is that software is mostly not about AI, that is, most of what we know about software applies to AI. In particular, many parts of the AI pipeline such as data collection and data cleansing are already familiar to software engineers. Even so, the data-driven aspect of AI poses some unique challenges such as dealing with the detection of bias in data, and dealing with incomplete and noisy data [61]. The second law is that AI software needs software engineers. Many of the activities of software engineering, such as system testing and maintenance, are activities that software engineers carry out routinely. The third law is that poor software engineering leads to poor AI, and the fourth law is the converse, that is, that better software engineering leads to better AI. These two laws are all about applying good SE practice when designing, developing and deploying AI software, taking into account the distinguishing features of AI systems. The fifth law is that even when using existing AI tools they still need to be adapted, using SE, to the application at hand. For example, it is generally useful to have tools to tune hyperparameters that could automatically improve predictive performance.

5.2 METHODS

The simplified ML workflow, shown in Figure 2, is the basis for DDD in the context of AI which is coupled with TEVV, as shown in Figure 4. A more detailed ML workflow (including, for

example, model requirements) was presented in [4], and a comprehensive workflow, which extends beyond ML to AI (including, for example, a review of AI ethics and the evaluation of metrics) was presented in [13].

An accepted distinction between software validation and verification presented in [8] is that validation addresses the question: “Are we building the right product?”, while verification addresses the question: “Are we building the product right?”.

It can be seen that for AI systems, validation is concerned with model evaluation guaranteeing that the conditions required for the deployment stage are satisfied, that is, ensuring that the predictions are accurate. We observe that while ML testing employs metrics such as accuracy, which take into account the overall performance of the system, conventional systems validate performance on individual inputs. On the other hand, the purpose of verification testing is to ensure that the product being developed meets the specified design according to the customer’s requirements.

A test method may be *static*, that is, it does not require program execution, or *dynamic*, which requires the execution of the program under development [2]. A prominent static method is formal verification, and its adaptation to AI [55] needs to take into account the distinguishing features of AI systems, and especially their non-deterministic behaviour. The promise of formal verification is to deliver provable guarantees of the correctness of the properties being tested. As a side effect, formal verification is also capable of providing comprehensible explanations raising the trustworthiness of the AI system; see Section 6.

Furthermore, with regards to test and evaluation, there is a need to take into account the distinguishing features of AI systems [11, 25]. To illustrate how these differences affect testing, assume a binary classification with classes A and B, with the output being the probability P that the input is A. If P is above a set threshold, then we choose A; otherwise we choose B. The choice of class can be compared to the ground truth of the input, if known, to indicate whether the AI system’s choice is correct or not. However, even if the choice of class is incorrect for one or more individual inputs, the model is not necessary invalidated, since the overall accuracy of a model is measured as the proportion of correct decisions made by the AI algorithm.

When we test a conventional algorithm against an input, then an error implies a “bug” in the code. For example, if the algorithm is meant to sort a set of numbers in ascending order, then there is only one correct result. Such a required method of testing software is called an *oracle test*; in ML the data is considered to be the oracle.

Several dynamic testing regimes, such as the following, have been adapted for testing AI systems [11]:

(1) *Differential testing*, where multiple implementations of the same algorithm are executed with the same data. Inconsistencies between the results indicate “bugs” in one or more of the implementations. Differential testing makes sense for deep neural networks (DNN), since the architecture and other aspects of a network can be defined so that different implementations are comparable.

(2) *Metamorphic testing*, where we make use of transformations, called metamorphic relations, that modify individual inputs to an ML model which should not affect their output. One example of a metamorphic relation is the rotation of an image, where we would expect the classification of the image to be unaffected by the rotation. Another simple

metamorphic relation involves the change of a time feature, so that if London is two hours behind Helsinki, then Helsinki is two hours ahead of London.

(3) *Fuzz testing*, where the operation of fuzzing involves providing invalid, unexpected or random data as inputs for testing. A form of fuzzing, called mutation testing, introduces small changes to inputs and may be viewed as a form of perturbation, although a mutation could, in principle, involve more significant changes than does a perturbation. Data can also be generated from scratch, for example, by a generative AI model.

The measure of *test adequacy* [11] addresses the problem of what are the criteria that constitute rigorous and thorough tests of software, so that the errors in the software are minimised. Now, assuming we run several tests on the software, then a test adequacy metric will evaluate whether the software has been sufficiently tested and identify any potential gaps in the testing. For example, in traditional software engineering, code coverage is a well established measure of the percentage of code that is executed under a particular testing regime. However, test adequacy criteria such as code coverage, focus on the internal structure of the code and are not suitable for AI systems. For a DNN it has been proposed to use neuron coverage, which is the ratio of the number of neurons in the network that have been activated to the total number of neurons in the network. However, there is no conclusive evidence that neuron coverage is sufficiently effective as a measure of test adequacy. Another type of test adequacy is combinatorial coverage, which considers the presence or absence of interactions between a small number of parameters, generally between two to six. The results in [12] show that, for AI systems, there is a correlation between test error and combinatorial coverage.

Two important aspects that test adequacy should measure are: (i) the similarity between the test and training data sets (low or high similarity should translate to low adequacy); and (ii) how closely is the test set aligned to the operational environment (low alignment should translate to low adequacy). The specification of test adequacy metrics for AI systems is an ongoing endeavour.

Although many researchers in SE for AI have focused their attention on software testing [40], there have also been studies on SE approaches for AI systems in the other major areas of the Software Engineering Body of Knowledge (SWEBOK) [33]. In particular, we also refer to data quality issues [61], which are important at all stages of the workflow of an AI system (see Figure 2), and the evaluation of the interaction of humans with the AI system [5], which is important for systems with partial autonomy, for AI products that respond to human input and provide outputs, and for tracking the performance of the AI system.

We conclude this section with a brief discussion of some important issues pertaining to SE for AI. First, benchmarking [60] is the provision of guidelines and best practices, containing two components: at least one data set and a reference implementation trained on the data set. Benchmarks benefit the community of AI developers, and allow for the comparison of performance of a variety of implementation strategies. Second, the availability of AI toolkits [17, 22] for general use is invaluable to the community of AI developers. Some of the benefits of using toolkits are: they lower the barriers to entry, they simplify the development of complex AI systems, they provide performance optimisation, and they may integrate with useful productivity tools such as visualisation design tools [49]. On the other hand, some downsides of using AI toolkits are: vendor lock-in, potential lack of development flexibility, possible security risks, and lack of access to the inner working of the code. On balance it would seem that the benefits outweigh the drawbacks, as without toolkits the barrier of entry

to the AI market would be too high for most developers. Finally, we mention the importance of open-source tools, which play a key role in the progress made in AI by providing open platforms for sharing and comparing results [64].

5.3 ARTEFACTS

The artefact output during the SE stage, resulting from the DDD-TEVV process, is an AI system, which is to be constrained by TAI metrics at the third stage described in Section 7. Although not a panacea, as mentioned above, from a pragmatic perspective there are quite a few AI toolkits available for general use, without which it would be very difficult to develop and maintain complex AI systems. Finally, we reemphasise that the life cycle is iterative and continuous at all of its stages, so we are unlikely to deploy an AI system until it is considered to be sufficiently trustworthy and safe, resolving the catalogued risks; see Section 6.

6 UNDERSTANDING TRUSTWORTHY AND SAFE AI

We now discuss the three main elements of the TAI life cycle, as shown in Figure 1, that is, the methodology, the methods employed and the artefacts produced, in the context of trustworthy and safe AI; see Figure 8.

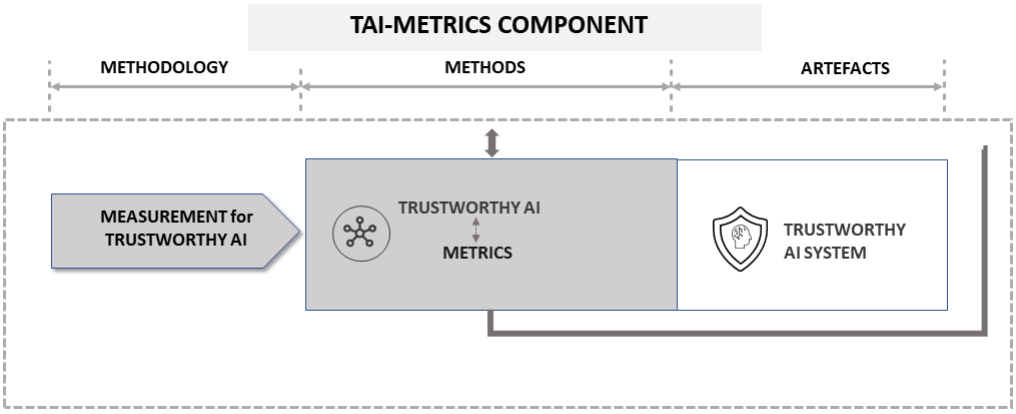

Figure 8: The TAI-metrics component of the TAI life cycle, shown in Figure 1.

6.1 METHODOLOGY

NPL, the UK's national metrology institute, is responsible for developing and maintaining the national primary measurement standards, and collaborates with metrology institutes around the world to maintain the international system of measurement. From a measurement perspective we emphasise uncertainty quantification (UQ), and its importance within the framework of TAI, to enhance transparency and trust in the outputs of AI systems [1]. Moreover, we stress that any measurement reported without an associated uncertainty is incomplete. We do not examine UQ any further here, assuming that it is embedded within the AI system, and refer the reader to [1] for more detail. We do, however, stipulate that measurement is a key component of the realisation of trustworthy and safe AI systems.

6.2 METHODS

Here, we will concentrate on the three TAI pillars: (i) technical, (ii) socio-technical and (iii) social, and their underlying principles, shown in Figure 5. These pillars are operationalised with metrics, which allow us to measure the trustworthiness of an AI system. Thus metrics are critical for transforming an AI system to a trustworthy and safe AI system (TAIS), which is what is needed to complete the TAI life cycle; metrics will be discussed in Section 7.

We now briefly discuss the various TAI principles [36, 37, 59]. Starting from the technical principles we have:

- Generalisability – The ability to make accurate decisions when confronted with new unseen data. This aspect is vital for the effectiveness of an AI system when it is being deployed. Generalisation often assumes that the distribution of new data is the same as that of the training data, an assumption which may be violated in practice. Overfitting should be avoided as far as possible, however, there are other challenges to do with data drift, when the distribution of the data changes in time.
- Reliability – The quality of a system performing consistently well as expected in normal circumstances; this aspect is related to the accuracy of the AI system.
- Resilience – The ability to recover or adapt to unanticipated failure events; this aspect is related to anomaly detection.
- Robustness – Already mentioned in the executive summary, the capability of high performance under a wide range of conditions, including unanticipated ones, such as noisy inputs.
- Security – Includes protocols designed to avoid, protect against, respond to, and recover from system failures and attacks. Security measures are implemented to safeguard AI systems from threats and vulnerabilities, which includes ensuring the system's resilience to attacks, its immunity to tampering, and its protection against abuse or misuse threats that can arise from any system components, including third-party ones.

We note that reliability, resilience and robustness (known collectively as R3) complement each other, and help us build TAI systems that can withstand both intentional (adversarial attacks) and unintentional (increases in the uncertainty of the input data) changes to the data processed by the systems' models.

Continuing to the socio-technical principles we have:

- Algorithmic fairness (or simply fairness) – Already mentioned in the executive summary, means that the AI system is free from bias and discrimination. In other words it is the manifestation of the social concept of fairness in AI systems. Complying with fairness criteria is widely accepted, and has applications in areas such as healthcare, criminal justice and financial services. We will discuss algorithmic fairness in more detail in Subsection 7.3.
- Autonomy and HITL – Already mentioned in the executive summary, where autonomy is the ability of an AI system to exhibit autonomous behaviour [59] and HITL controls the degree of human interaction with the AI system [63]. Figure 9 illustrates this point, with the two extremes being no autonomy and full autonomy, and allowing for partial

autonomy by varying the degree of human involvement in the decision-making process.

As an example, [58] investigate an autonomous parking feature. The researchers employed the intervention distance of drivers, using this feature as a measure of trust in autonomous parking. The smaller the distance the greater the trust. By collecting data on users' behaviour it is possible to ascertain the degree of partial autonomy that will be accepted and monitor how this changes in time as the performance of the AI improves.

- Explainability – Explainable means to make clear by describing in more detail or revealing relevant facts. So, for example, if a neural network predicts the temperature in your area, explainability would tell us *why* it arrives at a particular output. The goal of explainability is to help humans (for example, end-users) understand the decisions made by an AI system.
- Interpretability – Interpretable means to present in understandable terms. So, in our example above, interpretability would tell us *how* the neural network arrives at its output temperature. The goal of interpretability is to help humans (for example, developers) understand the inner working of the AI model when arriving at a decision.
- Privacy – This principle is about protecting individuals against unauthorised use of information that can directly or indirectly identify them. Privacy is compatible with data protection and other legislation that safeguards the collection, storage, and usage of personal data. (We note that privacy also applies to groups having well-defined common characteristics, but, for brevity, we will only refer to individuals here.) Although privacy concerns apply to all computer systems, privacy poses some unique challenges in AI systems due to its data-driven nature and, the likely complexity and evolving nature of its models. We will discuss privacy in more detail in Subsection 7.2.

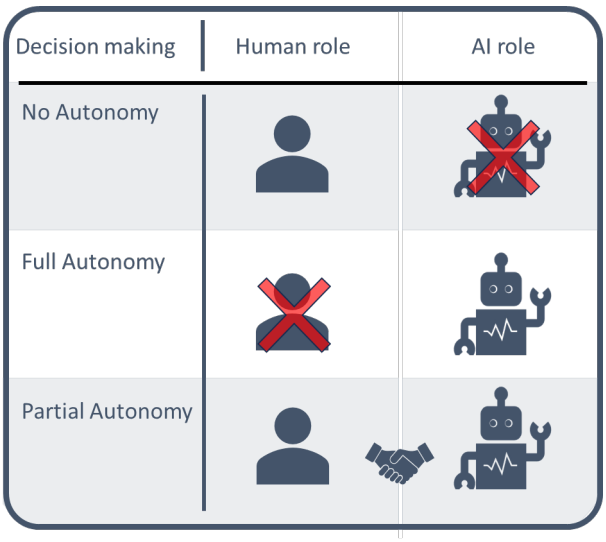

Figure 9: Levels of autonomy versus HITL.

Explainable and interpretable AI complement each other and, taken together, are known as XAI [15]; see Figure 10. We observe that XAI also addresses the *black-box* syndrome when an AI model is too complex to understand; this is especially prevalent in deep learning

models, which regularly have a large number of parameters that may be hard to interpret. We will discuss XAI in more detail in Section 7.1.

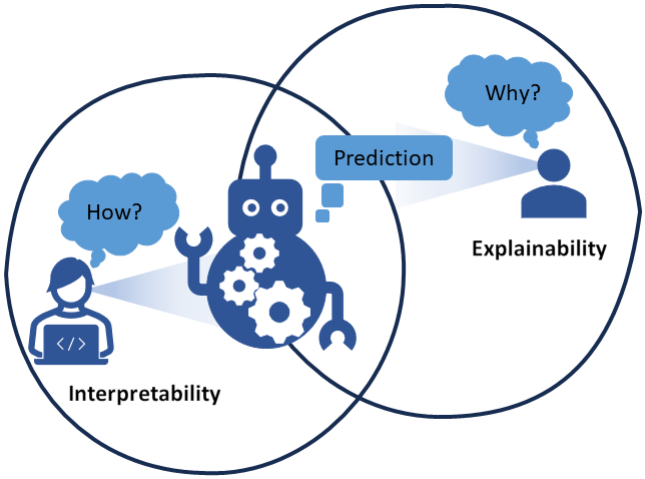

Figure 10: Interpretability versus explainability.

Ending with the social principles, that have ethical, legal and policy implications, we have:

- Accountability – Already mentioned in the executive summary, aims to ensure who is responsible for the outputs of an AI system. Auditability is derived from accountability, which requires an AI system to be reviewed, assessed, and audited, to identify problems in the trustworthiness of the system that need to be dealt with.
- Alignment – Also known as value alignment, ensures that the goals and behaviours of an AI system are compatible with human values and intentions [21]. Although alignment is primarily a social principle, it also has a socio-technical aspect, which is the challenge to incorporate human intentions and values into the AI life cycle.
- Beneficence – Beneficence means that AI systems should benefit humanity and the world we live in, by promoting human well-being and the environment, and respect basic human rights [20, 59]. Beneficience is grounded in AI ethics and the desire to embed these in all stages of the AI life cycle.
- Fairness – This is the social underpinning of algorithmic fairness, implying that the AI system is free of bias and discrimination, and promotes equitable outcomes and diversity. Fairness is also referred to as justice [20, 59], to emphasise the importance of adhering to an ethical and legal framework.
- Transparency – This is a broader principle than the scoio-technical principles of explainability and interpretability, whose aim is to ensure that humans have access to the AI system as a whole so that it can be understandable. It requires a level of disclosure of information pertaining to the AI system, covering the AI life cycle. Transparency promotes responsible and ethical design, development and deployment of AI systems. Nonetheless, there is a trade-off between transparency and privacy, as shown in Figure 11, where more transparency implies less privacy and vice versa.

We close this section with a brief discussion of generative AI (GAI), and the problems it poses to TAI. GAI is a form of AI that generates multimodal data, which is often synthetic

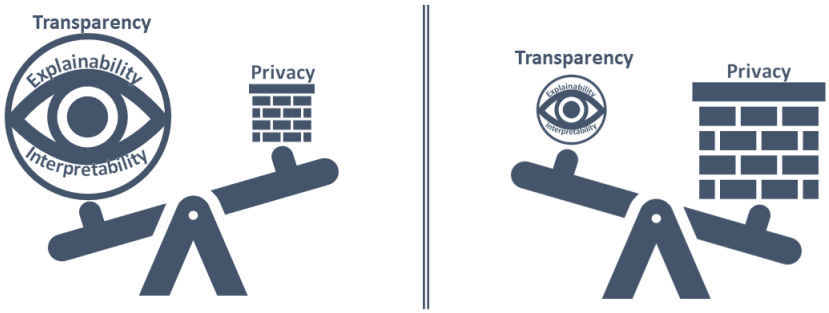

Figure 11: Transparency versus privacy trade-off.

[34]. Large language models (LLMs) are a prominent class of GAI neural network models, trained on very large corpora of text and mulitmodal data; LLMs are also known as foundation models [9]. LLMs such as ChatGPT [47] and Gemini [23] are a disruptive technology that could, amongst other things, transform search engines, lead to improved efficiency in content creation and programming, and, more generally, revolutionise automated question answering platforms.

GAI has many potential positive impacts, such as in drug discovery and materials science, and in improved efficiency and productivity (for example, in writing and programming). It does, however, also have potential negative impacts, which have already surfaced. In particular, we mention deepfakes (AI generated media) and fake news (false or misleading information spread as news) [27]. In many situations both deepfakes and fake news are a form of disinformation, that is, false information that is intentionally spread to mislead and cause harm. To mitigate against these serious threats, these GAI derivatives need to be detected and countered [24, 27]. One promising direction is to use automated fact checking combined with explainability [35], and another is to incorporate a secure way of embedding the provenance of the information into its metadata so that the provenance can be used to verify whether the information is true or not [27].

Given the importance of containing misinformation, noting that the meaning of trustworthy includes being truthful, there is an argument to be made for explicitly including truthfulness in our list of socio-technical principles; see [18].

6.3 Artefacts

The artefact output during the measurement stage, resulting from the process that specifies the relevant TAI metrics, examples of which will be given in Section 7, is a TAIS. This output is achieved by constraining the AI system by the TAI metrics that have been identified as relevant to the application at hand. As we have mentioned above, we cannot take it for granted that we will be able to satisfy all the relevant principles together; for example, see Figure 11 for the trade-off between transparency and privacy. Moreover, not all applications carry the same risks, which will affect the metrics that need to be applied to constrain the AI system; for example, spam filters and recommender systems are normally considered low risk applications.

7 METRICS FOR TRUSTWORTHY AND SAFE AI SYSTEMS

We now dig deeper, in Subsections 7.1, 7.2 and 7.3, into a selection of the socio-technical principles, namely, interpretability and explainability, privacy and algorithmic fairness (or simply fairness), respectively. In particular, we are interested in how metrics that enable us to operationalise the trustworthiness of an AI system with respect to these principles can be devised.

7.1 INTERPRETABILITY AND EXPLAINABILITY

Both interpretability and explainability are terms which are usually used to describe how understandable a specific model is [14]; see Subsection 6.2, where interpretability is defined as the *how* and explainability as the *why*.

Interpretability can describe a model, in which case the model will, in principle, be transparent, and the relationship between its input and output would be understandable to humans. It can also describe an algorithm, where in such a case, it will be easy to explain to a human in clear and understandable terms the steps of such an algorithm. In this sense interpretability is more comprehensive than explainability. On the other hand, transparency, which is a social principle of TAI, can be seen to encompass both interpretability and explainability, since it is a broader principle that is also related to the data not just the model; see Subsection 6.2. Moreover, interpretability and explainability do not describe the data directly but can shed light on its quality.

Explainability is often referred to as the ability to particularly explain the prediction (i.e. decision-making) process of an AI system. Explainability is further focused on a particular decision taken for a particular sample of the data. The explanation is then given to a human. As such, the receiving end of the explanation, who is a human, becomes part of the explanation process, since the quality of the explanation does not solely depend on the explanation itself, but also on how much of it was grasped by the receiver. Think of a particular explanation of a medical diagnosis given to the clinician, who most probably has an acceptable level of statistics and possibly mathematics. The very same explanation might not be understandable at all when given to the patient themself, who does not necessarily have any statistical, nor mathematical, background. In a nutshell, explainability focuses on why an algorithm made a specific decision as well as the justification of such a decision. Explainable AI represents one of the main linchpins on which a TAI systems should be founded. The quest for explainable AI systems has recently increased, since it is essential for sensitive applications, like self-driving cars and medical diagnosis.

For deep learning models, one seminal strategy of explainability, which is referred to as *local interpretable model-agnostic explanations* (LIME), was proposed in [54]. The main idea of LIME is to explain a black-box neural network by first perturbing the input, and then using the perturbations to establish a local linear model. This model then acts as a proxy which is simpler than the original black-box neural network, yet it explains the neural network within the neighbourhood of the input. LIME can also be used to identify input regions which are the most relevant for the neural network predictions. When dealing with image data, the fact that superpixels (a group of connected pixels with similar properties) must be identified a priori, makes LIME more difficult to deploy for some applications where such expert knowledge is not available.

We now discuss metrics for interpretability and explainability. Unlike accuracy, which is already well-established in the ML literature, there is no clear consensus yet on how to measure interpretability or explainability. We stress that explainability greatly depends on the human receiving the explanation, who remains the main judge of how good a particular explanation is.

Three main levels of evaluating interpretability were proposed in [14]:

(1) *Application-level evaluation*. This type can be applied to real-world tasks. It works by basing the explanation on a real experiment and then it being tested by an end-user. For example, think of a fracture detection software including an AI component which marks fractures in X-rays. An application-level evaluation would apply if a radiologist tests the fracture detection software in order to evaluate the system. A convenient baseline would be to have a human explaining the same decision.

(2) *Human-level evaluation*. This type is usually applied to simple tasks. It is a simplified version of the first evaluation type. The principal difference lies in the fact that experiments here do not have to be performed by domain experts. This leads to an, overall, easier (and cheaper) evaluation procedure. One example is to show a user different explanations from which they can choose the best explanation.

(3) *Functional-level evaluation*. This type does not require humans at all. It is based on a proxy task, and works best when the class of model used has already been evaluated by someone else in a human-level evaluation. For example, it might be well known that end-users can straightforwardly understand decision trees. In such case, a proxy for explanation quality can be the depth of the tree. Shorter trees would get a better explainability score. In order to guarantee an acceptable balance with accuracy, a constraint can be added which states that the predictive performance of the tree remains at an acceptable level (i.e. it should not decrease too much) compared to a larger tree.

7.2 Privacy

This subcomponent refers to personal privacy, data protection and security; see Subsection 6.2. It is ethically relevant to the concepts of autonomy, anonymity, confidentiality and control that are instrumental in guiding our choices for AI system design, development and deployment. Privacy is a major concern when deploying AI systems, for example, in the healthcare sector, especially when the data required for training AI models encode sensitive and confidential patient information. Any data breaches or misuse can result in substantial harm to patients, healthcare providers and software vendors. Most healthcare data sets are predominantly populated with patient information that should be mandatorily anonymised under the General Protection Data Regulation (GDPR). Privacy protection by an AI system can be scored based on its compliance with GDPR, documentation of privacy considerations (including consent of study subjects), strength of data security, data management plans (including a data life cycle), and anonymisation of personal sensitive information.

7.3 Fairness

This subcomponent [43] is complementary to data quality metrics [50]. Bias or exacerbation of disparities due to under-representation or incorrect representation within both training and validation data sets can have potentially unjust effects when deployed in the real world. In

the healthcare industry, this can have very adverse affects risking patient safety. Lack of appropriate representation can result in under-represented groups not receiving accurate treatment or any treatment at all. NIST's AI RMF defines three categories of bias that have to be considered and managed in a technical sense during the development of the model [57]: (i) systemic, (ii) computational and statistical, and (iii) human cognitive.

Each of these types of bias can occur in the absence of prejudice, partiality or discriminatory intent. Systemic biases can be present in AI data sets, organisational norms, practices and processes across an AI life cycle, and in the end-users. Computational and statistical biases are those present in the algorithmic processes and often stem from systemic errors due to incorrect or under-represented samples. Human cognitive biases relate to how an individual or a group perceives the AI output for decision-making or to fill in missing information.

Systemic biases can be minimised by ensuring diverse and relevant information is contained in the data sets used for training, validation and testing. Computational and statistical biases can be prevented by ensuring explainability within the AI system, allowing ease of debugging. Human cognitive biases can be minimised by providing adequate documentation and guidance for the understanding and operating of AI systems by end-users, with individual documentation directly based on the end-users' expertise and knowledge.

In ML, fairness refers to a procedure comprising an attempt to mitigate algorithmic bias, which, in this context, is bias that can potentially result from an ML-based prediction. The definition of fairness is disputed, and several definitions have been proposed. In general, when an ML model outputs a prediction which takes into account a sensitive attribute (for example, race, gender, and/or sexual orientation), this is considered unfair. Deciding which attributes are sensitive is another subjective topic, and it can differ according to the context, country and/or environment. We remark that a subset of sensitive attributes, called protected attributes, are those which are legally protected from being used in decision-making.

We close this section with a discussion of two types of metrics for fairness:

(1) *Disparate treatment*. One of the most common ways of measuring fairness is in terms of how limited disparate treatment is. More disparate treatment signifies more algorithmic bias and less fairness in the model. Disparate treatment takes place when an ML-based decision-making system outputs different predictions for groups of people who only differ in their sensitive attributes, i.e. they have equal (or similar) values of all the attributes apart from the sensitive attributes. For example, if two people (a male and a female), who have exactly the same salary, live in the same address, and are identical in all of their characteristics, apart from the gender, apply for a loan, and only the male gets his application accepted, then the model which took such a decision is considered one that is suffering from disparate treatment.

(2) *Disparate impact*. Here fairness is measured by the degree to which disparate treatment is mitigated. Disparate impact is present when an ML-based decision-making system systematically outputs predictions which adversely affects a group of people sharing, more frequently than another group of people, the same value for a sensitive attribute.

8 CONCLUDING REMARKS

We have presented a high-level life cycle for trustworthy and safe AI systems (TAIS); see Figure 1. The three layers of the life cycle are iterative and continuous, allowing transitions between the levels in either direction. Each layer consists of three components: methodology (stage), methods (processes) and artefacts (outputs). The definitive output from the life cycle is a TAIS artefact, which is an AI system constrained by the trustworthy and safe AI metrics defined at the third stage of the life cycle.

An AI workflow (extending the simplified ML workflow shown in the diagram in Figure 2) is a critical component in the development of an AI system, highlighting the distinguishing features of AI systems that need to be taken care of during the application of the TAI life cycle. These distinguishing features inform the ways in which the methodologies of risk management and SE need to change when applied to AI. They also frame an understanding of how metrics can be defined to address trustworthy and safe AI principles, allowing the transformation of an AI system into a TAI system.

We have provided sufficient detail on the RMF, DDD-TEVV, TAI, and metrics, in Sections 4, 5, 6 and 7, respectively, to understand what needs to be done to develop a TAIS. Although we have also indicated in some detail how this can be done, much of the detail has been left out of each layer due to space considerations and the many gaps in the current state of collective knowledge in this area that will need to filled in by AI researchers and practitioners. The next step is to apply the TAI life cycle to one or more use cases to trial its potential utility in practice.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the UK Department for Science, Innovation and Technology, which funded this work as part of the AI Standards Hub programme.

The authors would like to thank Peter Harris, Andrew Thompson, Sundeep Bhandari, Reva Schwarz and Elham Tabassi for helpful comments.

A APPENDIX: TAXONOMY MAPPING OF AI FRAMEWORKS

As discussed in Section 4 and Table 1, with the increased growth and popularity of the AI landscape, many institutes and organisations have proposed frameworks to promote and ensure trustworthy and safe AI. In this appendix, the aim is to provide a brief comparison of the principles and guidance recommended in four representative AI frameworks proposed by the CDEI, EU, NIST and OECD. The comparison is intended as a small-scale taxonomy mapping, and due to the difficulty of listing more AI frameworks as well as the terminologies employed in their work, only the high-level taxonomy from the four chosen representative frameworks are included here.

CDEI: In the UK, the CDEI has published an AI assurance roadmap [10] to improve the trustworthiness of AI. Since 2021, official language surrounding the CDEI has been concentrating on collaborativeness, particularly with industry. Key areas of focus include fairness, transparency, accountability, and public understanding.

EU: In Europe, AI regulation is the main focus. The EU's digital ethics strategy guidelines for TAI identifies seven key elements in TAI [28]. They are: (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, non-discrimination and fairness; (vi) environment and societal well-being; and (vii) accountability.

NIST: The NIST RMF [57] aims to minimise potential negative impacts in the development and deployment of AI systems. At the core of the NIST RMF are the principles of Govern, Map, Measure and Manage. The NIST RMF promotes TAI systems with characteristics including: (i) valid and reliable; (ii) safe; (iii) secure and resilient; (iv) explainable and interpretable; (v) privacy-enhanced and (vi) fair with harmful bias managed; and (vii) accountable and transparent.

OECD: The OECD also proposed essential guidance for responsible AI [45]. The guidance recommends the following five complementary key elements: (i) inclusive growth, sustainable development and well-being; (ii) human-centred values and fairness; (iii) transparency and explainability; (iv) robustness, security and safety; and (v) accountability. The principles recommend: (i) investing in AI research and development; (ii) fostering a digital ecosystem for AI; (iii) shaping an enabling policy environment for AI; (iv) building human capacity and preparing for labour market transformation; and (v) international co-operation for TAI.

A.1 High Level Principles

From the above description, a summary of the high level principles for TAI includes the following terms:

(a) Governance and accountability.

CDEI: Calls for clear accountability and oversight mechanisms; emphasises regular audits and evaluations.

EU: Proposes a regulatory framework to ensure accountability; calls for regulatory supervision and strict adherence to legal requirements.

NIST: Recommends clear lines of responsibility and accountability; encourages continuous monitoring and improvement.

OECD: Stresses the need for accountability and effective governance; encourages collaboration between governments, industry, and other stakeholders.

(b) Transparency and explainability.

CDEI: Advocates for clear communication of AI system capabilities and limitations; recommends explanations for significant decisions made by AI.

EU: Requires transparency and explainability for high-risk AI systems; proposes clear and accessible information for users.

NIST: Emphasises the importance of understandable AI systems; recommends transparent documentation and communication.

OECD: Encourages transparency and explainability for AI decisions; calls for accessible information for users.

(c) Fairness and bias mitigation.

CDEI: Calls for efforts to identify and mitigate bias in AI systems; recommends continuous monitoring for unintended consequences.

EU: Requires efforts to ensure fairness and avoid discrimination; proposes specific measures to address bias and discriminatory impacts.

NIST: Emphasises fairness and the mitigation of bias; recommends rigorous testing for bias in AI systems.

OECD: Stresses the importance of avoiding discrimination and bias in AI systems; recommends measures to address bias in training data.

(d) Human rights and societal impact.

CDEI: Emphasises the societal impact of AI on individuals and communities; recommends assessing and addressing social inequalities.

EU: Prioritises the protection of fundamental rights in the development and use of AI; proposes measures to address societal challenges and foster inclusivity.

NIST: Acknowledges the impact of AI on society and individuals; recommends considering societal and ethical norms.

OECD: Advocates for AI systems that respect human rights; calls for AI to contribute to sustainable development.

(e) Data privacy and security.

CDEI: Calls for responsible data practices and protection of privacy; emphasises the importance of data security.

EU: Prioritises the protection of personal data and privacy; proposes measures to enhance data security and privacy.

NIST: Emphasises the importance of privacy and security in AI systems; recommends protecting sensitive information.

OECD: Stresses the need to protect privacy and individual freedoms; recommends secure handling of data.

A.2 SUMMARY

Besides these high level taxonomies, many other terminologies are being used in AI frameworks. For instance, NIST has proposed taxonomies for technical and socio-technical attributes, and guiding principles contributing to AI trustworthiness. However, although similar, taxonomies may still have significant differences in their meaning. For example, “fair and bias managed” in NIST’s framework and “lawful and respectful of our nation’s values” in the US Executive Order [19]. It is also difficult to keep a uniform use of language across nations. Therefore, it may be better to use the taxonomies which are internationally recognised. For that purpose the definitions in the ISO standard [30] for AI concepts and terminology are suggested.

REFERENCES

[1] T. Adel, S. Bilson, M. Levene, and A. Thompson. “Trustworthy artificial intelligence in the context of metrology”. In: *Producing Artificial Intelligent Systems: The roles of Benchmarking, Standardisation and Certification*. Ed. by I. Ferreira. Computational Intelligence. Cham, Switzerland: Springer Nature, 2024.

[2] W.R. Adrion, M.A. Branstad, and J.C. Cherniavsky. “Validation, verification, and testing of computer software”. In: *ACM Computing Surveys* 14 (1982), pp. 159–192.

[3] F. Agbavor and H. Liang. “Artificial intelligence-enabled end-to-end detection and assessment of Alzheimer’s disease using voice”. In: *Brain Sciences* 13 (2023), Article 28, 13 pages.

[4] S. Amershi, A. Begel, C. Bird, R. DeLineand, et al. “Software engineering for machine learning: A case study”. In: *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. Montreal, 2019, pp. 291–300.

[5] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, et al. “Guidelines for human-AI interaction”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland, 2019, Paper 3, 13 pages.

[6] J. Banja. “How might artificial intelligence applications impact risk management?” In: *AMA Journal of Ethics* 22 (2020), pp. 945–951.

[7] A.J. Barnett, F.R. Schwartz, C. Tao, C. Chen, et al. “A case-based interpretable deep learning model for classification of mass lesions in digital mammography”. In: *Nature Machine Intelligence* 3 (2021), pp. 1061–1070.

[8] B.W. Boehm. “Verifying and validating software requirements and design specifications”. In: *IEEE Software* 1 (1984), pp. 75–88.

[9] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, et al. “On the opportunities and risks of foundation models”. In: *Machine Learning Archive* arXiv:2108.07258 [cs.LG] (2022).

[10] CDEI. *The roadmap to an effective AI assurance ecosystem*. Tech. rep. See https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem. London, UK: Centre for Data Ethics and Innovation (CDEI), Dec. 2021.

[11] J. Chandrasekaran, T. Cody, N. McCarthy, E. Lanus, et al. “Test & evaluation best practices for machine learning-enabled systems”. In: *Software Engineeering Archive* arXiv:2310.06800 [cs.SE] (2023).

[12] T. Cody, E. Lanus, D.D. Doyle, and L. Freeman. “Systematic training and testing for machine learning using combinatorial interaction testing”. In: *Proceedings of the IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. Valencia, Spain, 2022, pp. 102–109.

[13] D. De Silva and D. Alahakoon. “An artificial intelligence life cycle: From conception to production”. In: *Patterns* 3 (2020). See https://doi.org/10.1016/j.patter.2022.100489, 13 pages.

[14] F. Doshi-Velez and B. Kim. “Towards a rigorous science of interpretable machine learning”. In: *Machine Learning Archive* arXiv:1702.08608 [stat.ML] (2017).

[15] R. Dwivedi, D. Dave, H. Naik, S. Singhal, et al. “Explainable AI (XAI): Core ideas, techniques, and solutions”. In: *ACM Computing Surveys* 55 (2023), Article 194, 33 pages.

[16] D. Elgesem. “The AI act and the risks posed by generative AI models”. In: *Proceedings of the symposium of the Norwegian AI Society (NAIS), CEUR Workshop Proceedings, Volume 3431*. 8 pages. Bergen, Norway, 2023.

[17] B.J. Erickson, P. Korfiatis, Z. Akkus, T. Kline, et al. “Toolkits and libraries for deep learning”. In: *Journal Of Digital Imaging* 30 (2017), pp. 400–405.
[18] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, et al. “Truthful AI: Developing and governing AI that does not lie”. In: *Computers and Society Archive* arXiv:2110.06674 [cs.CY] (2021).
[19] Executive Order 13960. *Promoting the Use of trustworthy artificial intelligence in the federal government*. Executive Office of the President, Washington DC, https://www.federalregister.gov/d/2020-27065. Dec. 2020.
[20] L. Floridi and J. Cowls. “A unified framework of five rrinciples for AI in society”. In: *Harvard Data Science Review (HDSR)* 1 (2019). https://doi.org/10.1162/99608f92.8cd550d1, 14 pages.
[21] I. Gabriel. “Artificial Intelligence, values, and alignment”. In: *Minds and Machines* 30 (2020), pp. 411–437.
[22] M. Gheisari, F. Ebrahimzadeh, M. Rahimi, M. Mohamadtaghi, et al. “Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey”. In: *CAAI Transactions on Intelligence Technology* 8 (2023), pp. 581–606.
[23] Google. *Gemini*. https://gemini.google.com. Accessed May 2024. 2024.
[24] A.A. Hamed, M. Zachara-Szymanska, and X.Wu. “Safeguarding authenticity for mitigating the harms of generative AI: Issues, research agenda, and policies for detection, fact-checking, and ethical AI”. In: *iScience* 27 (2024). https://doi.org/10.1016/j.isci.2024.108782, 6 pages.
[25] D.J. Hand and S. Khan. “Validating and verifying AI Systems”. In: *Patterns* 1 (2020), Article 100037, 3 pages.
[26] N.A.J. Hastings. “Physical Asset Management: With an Introduction to the ISO 55000 Series of Standards”. In: Second. Cham, Switzerland: Springer International Publishing, 2015. Chap. 15 Risk Analysis and Risk Management, pp. 249–270.
[27] T.C. Helmus. *Artificial Intelligence, deepfakes, and disinformation A Primer*. Tech. rep. See www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf. Santa Monica, Ca.: RAND Corporation, July 2022, 24 pages.
[28] High-Level Expert Group on Artificial Intelligence. *Ethics guidelines for trustworthy AI*. Tech. rep. See https://doi.org/10.2759/346720. Brussels: European Commission, Directorate-General for Communications Networks, Content, and Technology, Apr. 2019, 41 pages.
[29] ISO/IEC. *Information technology — Artificial intelligence - AI system life cycle processes*. International Standard ISO/IEC 5338:2023(E). 39 pages. Geneva, CH: International Organization For Standardization, 2023.
[30] ISO/IEC. *Information technology — Artificial intelligence - Artificial intelligence concepts and terminology*. International Standard ISO/IEC FDIS 22989:2022(E). 60 pages. Geneva, CH: International Organization For Standardization, 2023.
[31] ISO/IEC. *Information technology — Artificial intelligence — Guidance on risk management*. Standard ISO/IEC 23894:2023(E). 26 pages. Geneva, CH: International Organization For Standardization, 2023.
[32] ISO/IEC. *Information technology — Artificial intelligence — Management System*. International Standard ISO/IEC FDIS 42001:2023(E). 51 pages. Geneva, CH: International Organization For Standardization, 2023.
[33] ISO/IEC. *Software Engineering — Guide to the Software Engineering Body of Knowledge (SWEBOK)*. Technical Report ISO/IEC TR 19759:2015. 336 pages. Geneva, CH: International Organization For Standardization, 2016.

[34] M. Jovanović and M. Campbell. “Generative artificial intelligence: Trends and prospects”. In: *IEEE Computer* 55 (2022), pp. 107–112.
[35] N. Kotonya and F. Toni. “Explainable automated fact-checking: A survey”. In: *Proceedings of the 28th International Conference on Computational Linguistics (ACL)*. Barcelona, 2020, pp. 5430–5443.
[36] M. Levene and J. Wooldridge. *Certification of machine learning applications in the context of trustworthy AI with reference to the standardisation of AI systems*. NPL Report MS 45. See https://doi.org/10.47120/npl.MS45. National Physical Laboratory (NPL), Mar. 2023.
[37] B. Li, P. Qi, B. Liu, S. Di, et al. “Trustworthy AI: From principles to practices”. In: *ACM Computing Surveys* 55 (2023), Article 177, 46 pages.
[38] Q. Liu, Z. Chen, H. Li, M. Huang, et al. “Modular end-to-end automatic speech recognition framework for acoustic-to-word model”. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (2020), pp. 2174–2183.
[39] B. López. *Case-Based Reasoning: A Concise Introduction*. Synthesis Lectures on Artificial Intellgence and Machine Learning. San Rafael, CA: Morgan & Claypool Publishers, 2013.
[40] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, et al. “Software engineering for AI-based systems: A Survey”. In: *ACM Transactions on Software Engineering and Methodology* 31 (2022), Article 37e, 59 pages.
[41] J. Matthews. *How should we regulate AI?* Public Policy Report. 22 pages. IEEE-USA, Aug. 2023.
[42] T. Menzies. “The Five Laws of SE for AI”. In: *IEEE Software* 37 (2020), pp. 81–85.
[43] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, et al. “Algorithmic fairness: Choices, assumptions, and definitions”. In: *Annual Review of Statistics amd its Application* 8 (2021), pp. 141–163.
[44] R.J. Neuwirth. “Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA)”. In: *Computer Law & Security Review* 48 (2023). See https://doi.org/10.1016/j.clsr.2023.105798, 14 pages.
[45] OECD. *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*. Tech. rep. 312. See https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm. Organisation for Economic Co-operation and Development (OECD), June 2021.
[46] OECD.AI Policy Obsevatory. *OECD AI Principles overview*. https://oecd.ai/en/ai-principles. 2024.
[47] OpenAI. *ChatGPT*. https://chat.openai.com. Accessed May 2024. 2024.
[48] A. Pal and Y. Rathi. “A review and experimental evaluation of deep learning methods for MRI reconstruction”. In: *Journal of Machine Learning for Biomedical Imaging* 1 (2022). See https://doi.org/10.59275/j.melba.2022-3g12, 50 pages.
[49] N. Pezzotti, T. Höllt, J. van Gemert, B.P.F. Lelieveldt, et al. “DeepEyes: Progressive visual analytics for designing deep neural networks”. In: *IEEE Transactions on Computer Graphics* 24 (2018), pp. 98–108.
[50] E. Pitoura. “Social-minded measures of data quality: Fairness, diversity, and lack of bias”. In: *ACM Journal of Data and Information Quality* 12 (2020), Article 12, 8 pages.
[51] P. Rajpurkar and M.P. Lungren. “The current and future state of AI interpretation of medical images”. In: *The New England Journal of Medicine* 388 (2023), pp. 1981–1990.

[52] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, et al. “Explainable, trustworthy, and ethical machine learning for healthcare: A survey”. In: *Computers in Biology and Medicine* 149 (2022), Article 106043, 23 pages.

[53] S. Reddy, W. Rogers, V. P. Makinen, E. Coiera, et al. “Evaluation framework to guide implementation of AI systems into healthcare settings”. In: *BMJ Health Care Inform* 28 (2021). See http://dx.doi.org/10.1136/bmjhci-2021-100444, 7 pages.

[54] M. Ribeiro, S. Singh, and C. Guestrin. “Why should I trust you? Explaining the predictions of any classifier”. In: *Proceedigns of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, 2016, pp. 1135–1144.

[55] S.A. Seshia, D. Sadigh, and S.S. Sastry. “Towards verified artificial intelligence”. In: *Communications of the ACM* 65 (2022), pp. 46–55.

[56] K. Si, Y. Xue, X. Yu, X. Zhu, et al. “Fully end-to-end deep-learning-based diagnosis of pancreatic tumors”. In: *Theranostics* 11 (2021), pp. 1982–1990.

[57] E. Tabassi. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. See https://doi.org/10.6028/NIST.AI.100-1. Gaithersburg, MD: NIST Trustworthy, Responsible AI, National Institute of Standards, and Technology, Jan. 2023, 48 pages.

[58] N.L. Tenhundfeld, E.J. de Visser, K.S. Haring, A.J. Ries, et al. “Calibrating trust in automation through familiarity with the autoparking feature of a Tesla model X”. In: *Journal of Cognitive Engineering and Decision Making* 14 (2023), pp. 279–294.

[59] S. Thiebes, S. Lins, and A. Sunyaev. “Trustworthy artificial intelligence”. In: *Electronic Market* 31 (2021), pp. 447–464.

[60] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey. “Scientific machine learning benchmarks”. In: *Nature Review Physics* 4 (2022), pp. 413–420.

[61] S.E. Whang, Y. Roh, H. Song, and J.-G. Lee. “Data collection and quality challenges in deep learning: a data-centric AI perspective”. In: *The VLDB Journal* 32 (2023), pp. 791–813.

[62] T.M. Willams. “Using a risk register to integrate risk management in project definition”. In: *Intemational Journal of Project Management* 12 (1994), pp. 17–22.

[63] X. Wu, L. Xiao, Y. Sun, J. Zhang, et al. “A survey of human-in-the-loop for machine learning”. In: *Future Generation Computer Systems* 135 (2022), pp. 364–381.

[64] Z. Yang, C. Wang, J. Shi, T. Hoang, et al. “What do users ask in open-source AI repositories? An empirical study of GitHub issues”. In: *Proceedings of the IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*. Melbourne, Australia, 2023, pp. 79–91.

[65] R. Zhao, Y. Zhang, B. Yaman, M.P. Lungren, et al. “End-to-end AI-based MRI reconstruction and lesion detection pipeline for evaluation of deep learning image reconstruction”. In: *Computer Vision and Pattern Recognition Archive* arXiv:2109.11524 [cs.CV] (2021).