NPL
National Physical Laboratory

NPL REPORT MS 45

CERTIFICATION OF MACHINE LEARNING APPLICATIONS IN THE CONTEXT OF TRUSTWORTHY AI WITH REFERENCE TO THE STANDARDISATION OF AI SYSTEMS

MARK LEVENE, JENNY WOOLDRIDGE

MARCH 2023

Certification of Machine Learning Applications in the Context of Trustworthy AI with Reference to the Standardisation of AI Systems

Mark Levene and Jenny Wooldridge
Data Science Department

ISSN 1754-2960

https://doi.org/10.47120/npl.MS45

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

This work was funded by the UK Government's Department for Science, Innovation & Technology.

Approved on behalf of NPLML by
Louise Wright, Head Of Digital Metrology, Data Science Department.

CONTENTS

ACRONYMS

AI	Artificial Intelligence
AIA	Artificial Intelligence Act
BNN	Bayesian Neural Network
CAV	Connected Autonomous Vehicles
CDEI	Centre for Data Ethics and Innovation
ChatGPT	Chat Generative Pre-trained Transformer
DNN	Deep Neural Network
EMN Mathmet	European Metrology Network for Mathematics and Statistics in Metrology
ETSI	European Telecommunications Standards Institute
GPAIS	General Purpose Artificial Intelligence Systems
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
ITU-T	International Telecommunication Union Telecommunication Standardisation Sector
LLM	Large Language Model
ML	Machine Learning
NN	Neural Network
NPL	National Physical Laboratory
NQI	National Quality Infrastructure
SRA	Strategic Research Agenda
UQ	Uncertainty Quantification
XAI	Explainable Artificial Intelligence

ABSTRACT

Artificial intelligence (AI) and its subset machine learning (ML), which focuses on learning tasks from data, are some of the most disruptive emergent technologies. AI standards inform organisations how to develop and manage their AI systems and are emerging to satisfy the increasing demand from industry and society for the safe adoption of AI and ML technologies. However, AI systems must be trustworthy in the sense that they can be relied upon to make responsible decisions. Consequently, trustworthy AI is a collection of principles that encourages responsible development, use and deployment of AI systems, and can be viewed as a framework for managing risk in AI systems.

The National Physical Laboratory (NPL) is one of the four institutions responsible for delivering the UK's national quality infrastructure (NQI), in which standards and certification play key roles. In this context we review research in NPL on trustworthy AI, emphasising the importance of uncertainty quantification (UQ) in enhancing the transparency and trust in results output from AI systems.

We review the landscape of AI standards and certification and emphasise their role in the context of trustworthy AI. Third-party certification is a key service in building trust in AI and ML systems and supporting their operationalisation. We argue that certification should assess conformity to AI standards and characteristics of trustworthy AI, and, in addition, should be able to carry out conformity testing and evaluation of the components of an AI system. As a case study we look at ChatGPT, a large AI system which is attracting a lot of attention, and investigate its potential conformity to the principles of trustworthy AI.

1 INTRODUCTION

Artificial Intelligence (AI) and machine learning (ML) are rapidly evolving technologies that have revolutionised the way we interact with computers and the world around us. AI is the theory and development of computer systems that are able to perform tasks that normally require human intelligence, of which machine learning is a subset, focussing on AI methods that are able to learn and adapt. AI includes symbolic computation, such as expert systems, which are not a part of ML, whereas ML builds statistical models of data that may be used for classification and prediction tasks to aid decision-making. ML models are generally grouped into supervised and unsupervised learning, with the former learning how to make predictions and classifications by training and testing a model from data which has already been labelled by humans, and the latter identifying clusters and patterns within unlabelled data with no additional human input. Self-supervised learning is a third type of machine learning, where data is labelled automatically by the machine, as a precursor for a supervised task. It alleviates the need for human-annotated labels that are generally not available in abundance and costly to produce. Here we will mainly focus on ML, but will often refer to AI as the more general technology.

AI standards, and in particular the recent ISO/IEC DIS (Draft International Standard) 42001 Artificial Intelligence Management System [1], which is still under development as of March 2023, inform organisations how to develop and manage their AI systems. Standards, such as ISO/IEC 23053, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) [2], in addition, foster a common terminology for AI concepts and provide a basis for clearly describing the design, engineering and usage of AI systems. In this report we mainly concentrate on the certification of AI systems, which is a third-party service whose aim is to assess the conformity of an AI system to some set of criteria. We argue that certification should take place along three dimensions:

(i) against standards such as ISO/IEC DIS 42001,
(ii) against criteria pertaining to the “trustworthiness” of the system (discussed in detail below), and/or
(iii) conformity testing and evaluation of the AI components of the system.

Whilst the concepts surrounding certification of AI discussed here are generalisable to all types of learning, we will mainly focus on supervised learning but also consider self-supervised learning when discussing the much talked about ChatGPT [3], which is to a large degree trained in a self-supervised manner.

In the rest of this report, we describe the current landscape of certification to documentary standards for AI products and processes, and the challenges that AI poses to traditional certification methods. Since the AI standards landscape is still “work in progress”, this report has broadened the scope of certification to include trustworthiness characteristics and testing/evaluation of the AI components of the system. We also present a case study on the recent large uptake of large language models (LLMs, or more generally foundation models, which may be trained on multimodal data rather than text as are LLMs), in particular ChatGPT, discussing the risks involved with the use of such technologies should they be unregulated and/or uncertified. We emphasise that certification of AI systems will require joint efforts between organisations having expertise in trustworthy AI, to complement and inform progress in current AI standardisation. Finally, uncertainty quantification (UQ) plays a key role in the evaluation of AI systems. UQ helps generate trust in AI systems, and as such should be a critical component of AI model validation and certification.

2 TRUSTWORTHY AI

The word trustworthy means “able to be relied upon as honest and truthful”. In the context of AI, the requirement that a model is trustworthy also calls for the application of traditionally anthropomorphic characteristics such as morality and ethics to IT systems. In the sections below we introduce the concepts behind trustworthy AI and current research activity at NPL on this topic.

2.1 THE PILLARS OF TRUSTWORTHY AI

As AI and ML models become more prevalent in our daily lives, it is increasingly important to ensure that such models are “trustworthy” and can be relied upon to make responsible decisions. A major theme in AI is for organisations both developing and using AI systems to have confidence in the technology. Thus, trustworthy AI, or more specifically trustworthy ML, is crucial for responsible development and use of the technology to take place. Trustworthy AI can be viewed as a framework for managing risk in an AI system [4], and can be divided into three general themes, each with its underlying principles, topics and characteristics (see Figure 1).

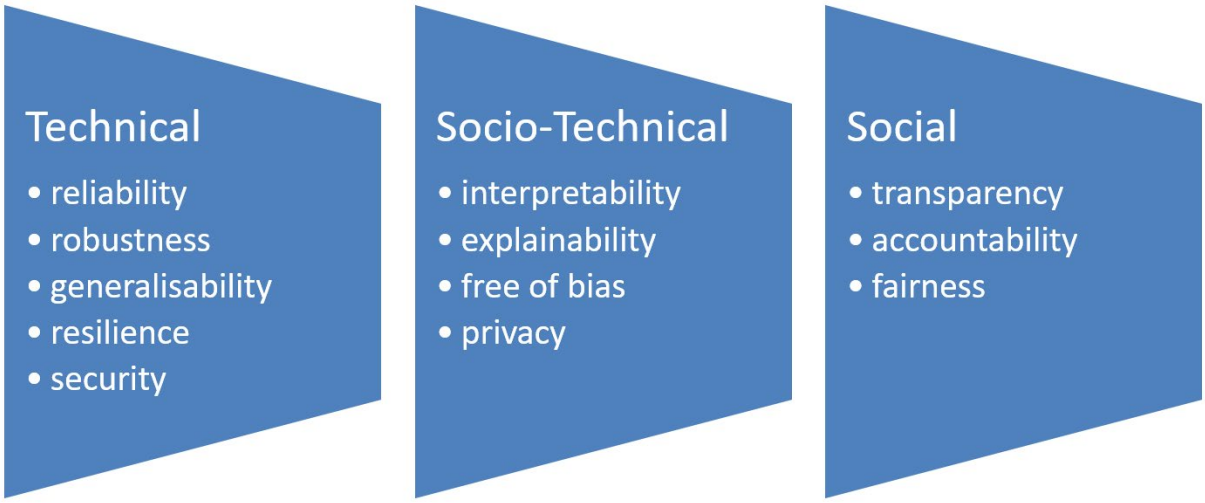

Figure 1. The three pillars of trustworthy AI.

Fairness is one of the fundamental requirements of trustworthy AI, with the intention of identifying underlying bias in the programmatic decision-making of such systems. Therefore, the focus of some of the current research on trustworthy AI centres on ways of quantifying these characteristics through the development of fairness scores and bias indices. In [5] a bias index is proposed, which measures an aggregate score across a variety of fairness metrics evaluated on the AI system. On an individual level, fairness implies that if two individuals differ only in protected attributes, such as gender or race, their outcomes should be similar, otherwise there is individual discrimination. This can be extended to groups by dividing the data set based on protected attributes, and testing whether the outcomes are similar across the groups. The bias index essentially evaluates whether the difference between the probabilities of the predicted labels in a classification problem of the privileged and underprivileged groups is above a given threshold. Such a score could be used to operationalise fairness certification.

The reliability of a system is the quality of performing consistently well as expected in normal circumstances, while robustness is the capability of high performance under a wide range of conditions (in ML we often consider robustness to noise perturbations in the input). Resilience, on the other hand, is the ability to recover or adapt to unanticipated failure

events. Trustworthy AI also requires the model to be reliable, robust and resilient in the face of both intentional (adversarial attacks) and non-intentional (increases in input uncertainty) changes to the data processed by the model. With the former, the model inputs are maliciously varied to attempt to fool the model, and recent developments in AI research have produced an arms-race in methodologies that aim to ensure robustness in the face of such attacks [6]. This has prompted researchers to consider the need for the certification of robustness, requiring measured proof that the model yields reliable and expected outputs when subject to threats on the inputs [7]. It is also critical to certify that the training and test data are similar enough in their distribution of data items, to ensure that the prediction model is relevant to the data on which it is being applied, i.e. that there is no unauthorised extension to the model application beyond the scope of its original intended use case.

Explainability and interpretability are two other key considerations in trustworthy AI. Ideally, we would like our ML models to be both "interpretable" (to present in understandable terms) and "explainable" (to make clear by describing in more detail or revealing relevant facts). As an example, if a neural network (NN) predicts the future temperature at a specific location, we would like to understand "how" the NN arrives at its output (interpretability), and to get an explanation of "why" it arrives at a particular output (explainability). Interpretability may help in debugging ML software by tracing the manner in which an answer was produced, while explainability will increase our trust in the AI system.

Explainability also promotes fairness and dissuades from discrimination by increasing the transparency of the model. Transparency relates both to the details of the AI model and to the specifics of the systems and processes of the organisations developing and/or operating the AI system, including the methodology used to collect and process data used within the model [8]. The goal of explainable AI (also known as XAI) is to help humans understand decisions made by an AI system, addressing the “black box syndrome”, when the model is too complex. NN architectures pose specific challenges to model explainability. They are constructed with nodes (known as neurons) on which we calculate basic computational operations that are propagated through links connecting the nodes. Thus in theory the internal states of the network can be traced and reproduced at all points within it. However, such openness and visibility does not in itself enable any kind of interpretability of what the model is doing, due to the sheer number of operations that make it impossible for a human to understand the process as a whole. This is especially true in what are known as deep neural networks (DNNs), which stack up multiples layers of neurons between the inputs and outputs of the NN. Again, certification of AI systems can help with explainability, not in terms of requiring exact reproducibility under ideally defined conditions, but in ensuring that the implementation of the ML models meet some minimum performance requirements in a statistically significant fashion [9].

Uncertainty quantification also plays an important, but often overlooked, role in AI models. By quantifying the uncertainty of trustworthy characteristics we are able to be transparent about the limitations of ML results, which provides crucial information on the reliability, robustness and explainability of results [10]. There are uncertainties associated with each step of the machine learning cycle, from the data preparation (aleatoric uncertainty) and model construction (including the propagation of uncertainty within the computational operations performed within the model) to model evaluation (epistemic uncertainty). From a user perspective, communicating the uncertainties can provide a crucial layer enhancing transparency and trust in the AI system. However, most ML models do not generate a confidence measure of the uncertainties on the result, but instead generate point estimates.

Classification models may output probability values, for example obtained in the top layer of a NN, and these are often interpreted as model confidence. However, these probabilities do not represent the measurement uncertainty in the metrology sense, since on their own they are point estimates and thus do not characterise the dispersion of values of the output,

unless we have the additional knowledge of their underlying probability distribution [11]. Still, combining calibration values with predicted probabilities is important in evaluating a ML model, and can improve users' trust in the AI system [12]. We emphasise the metrology aspect here that it should be possible to trace data back to the source of measurement, as the uncertainty will propagate through the system from the input to the output. There are generally two types of methods for UQ in ML, Bayesian and frequentist, depending on how the uncertainties are obtained. Bayesian methods for measuring uncertainty, such as Bayesian neural networks (BNNs), are designed to produce uncertainty estimates along with their predictions, while in frequentist methods, such as ensemble methods, uncertainty is measured by retraining the network several times and averaging the results [13].

2.2 RESEARCH ON TRUSTWORTHY AI AT NPL

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. As part of the research in the Data Science Department at NPL, we work on trustworthy ML with a particular focus on explainability and UQ in the context of metrology [14]. It is important to note that in metrology no measurement is complete without a statement of the associated uncertainty, i.e. the degree of belief about the dispersion of the measured value. As previously mentioned, there are two types of uncertainties that are generally considered in ML: aleatoric uncertainty (uncertainty about the data), and epistemic uncertainty (uncertainty about the model) [15]. From a metrology perspective we are also concerned with tracing the uncertainty back to the source of measurement before it was input to the system, and quantifying the combined uncertainty of the output.

Much of the research at NPL is applied with the aim of delivering impact in areas of national and international challenge such as energy and environment, life sciences and health, and connected and autonomous vehicles (CAV). In the context of AI and ML we do applied research in these areas in the Data Science department. Some applications we are looking at include assured autonomy of marine vessels, the effect of weather on sensor performance for autonomous cars, medical imaging, digital healthcare data curation and management, and applications in advanced manufacturing. It is also worth mentioning that we are also at the forefront of digital metrology [16], for example the specification of digital calibration certificates [17].

NPL has been instrumental in drafting the agenda for machine learning and artificial intelligence as part of the Strategic Research Agenda (SRA) for the European Metrology Network for Mathematics and Statistics in Metrology (EMN Mathmet); the draft is open for comment until April 2023 [18]. The SRA details mathematical and statistical issues that contribute to the trustworthiness of a ML prediction. It also emphasises the importance of a quality framework for guiding the choice of a ML model, considering the quality and provenance of the data, and providing verification and validation of the ML algorithms and software used. The framework supports the reproducibility of results and auditability of ML models in metrology applications. It also highlights the importance of the specification of a standard interface for benchmarking, validation and certification of ML models. Finally, it presents a roadmap for the metrology community to meet the challenges of ML and AI.

3 STANDARDS AND CERTIFICATION LANDSCAPE

The UK's National Quality Infrastructure (NQI) [19] is instrumental in ensuring that businesses and consumers have confidence in products and services. The certification of

products and services guarantees conformity to national and/or international standards, and enables businesses to export their products globally. The five main components of the NQI are shown in Figure 2.

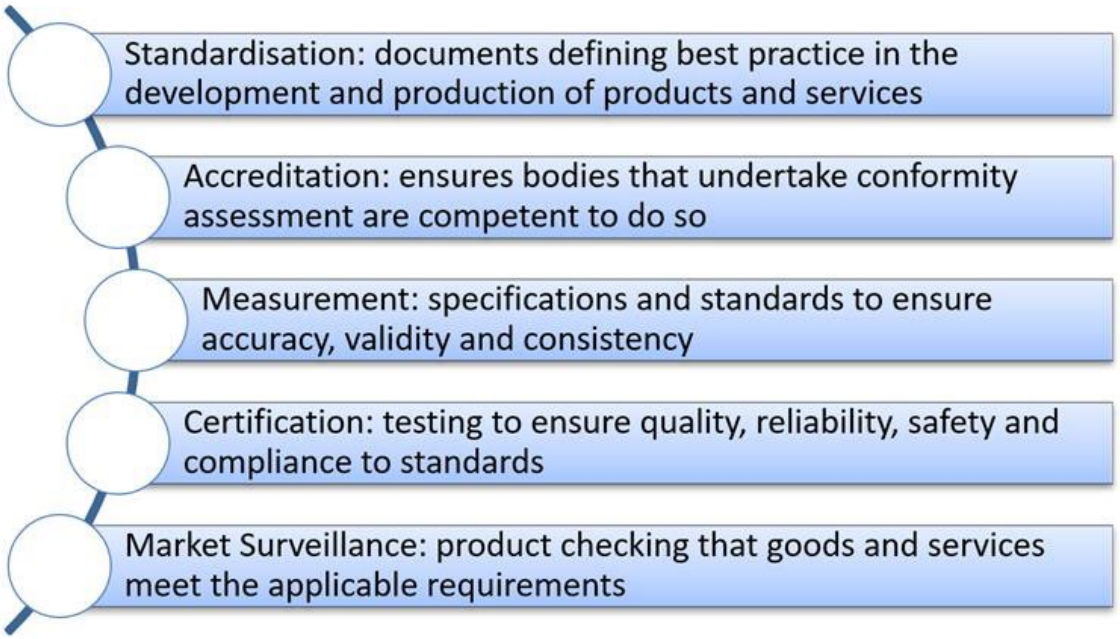

Figure 2. Components within the UK National Quality Infrastructure

The application of the NQI framework to the development and marketing of physical goods and traditional software products is well established, and there is a large evidence base to show how standards and certification reduces transaction costs and enables markets to function more efficiently [20]. NPL is one of the four main institutions that are responsible for delivering the NQI, accountable for maintaining and disseminating primary measurement standards to ensure the accuracy and consistency of measurement. These top-level national standards (which are compared internationally with those of other national measurement institutes around the world through a framework known as the Mutual Recognition Arrangement) are disseminated through a large network of calibration and testing laboratories, which have been accredited to undertake conformity assessment to standards.

A primary role of certification in the context of AI is to reduce information asymmetries between the suppliers and users of AI systems by accurately characterising the object of certification, justifying the requirement for an independent third party to carry out the conformity assessment [21]. It is important that certification is adopted substantively and not just symbolically, which can be achieved by offering the right incentives for conformity. The level of certification recommended will depend on the level of risk of the application under consideration.

The disruptive nature of AI and ML products, however, presents a series of challenges to the existing systems and processes of the quality infrastructure. Firstly, we should distinguish between product and process certification, with the former relating to demonstrating compliance at the delivery level, and the latter covering the management systems used to deliver the product [22]. The certification of management and quality systems is perhaps easier to address, as there is a degree of familiarity and overlap to existing standards such as ISO 9001 [23] or ISO 27001 [24]. However, in applying such standards to the management of AI systems, care must also be taken to address ethical issues (behavioural guidelines accepted by the community) in ways that are perhaps not applicable to non-AI processes [25]. Moreover, new guidelines are required to develop novel quality metrics that

describe non-traditional considerations such as social, ethical and diversity dimensions [26]. This has led to the suggestion of borrowing concepts from the governance of value chains and commodities within sustainability certification, which also have ethical implications and are often similarly difficult to quantify [27]. Conformity to standards by definition requires that all the desirable properties of the system can be formalised, and as such verification of AI systems is limited to aspects that can be quantified, or for which goal definitions can be explicitly defined. Nonetheless, it is important that social considerations and design to account for ethical consequences take place before the deployment of these technologies and not after, as has been the case for ChatGPT and other LLMs and foundation models that have been released concurrently. It is crucial that such foundation models be tested against the trustworthy AI characteristics, which can be viewed as providing a framework for managing risk in an AI system [4, 28].

In addition, formal verification of ML models through certification by a trusted third party can build confidence for the end-user for the overall quality and safe usage of the technology. One aspect of ML algorithms which contrasts with traditional software products is the role of data within the model itself. Whilst ML models are inherently probabilistic, transforming inputs from the real world into outputs (say, in a classification task) that can be qualified with associated probabilities. The values of those output probabilities are determined from the data used to train the model. This raises a number of verification challenges which must be addressed. ML models are generally better in terms of accuracy (minimising the difference between the predicted and actual classification labels) when the model is trained on larger data sets, however it is not guaranteed that the trustworthiness of a model improves with the scaling of a training data set in the same way [29]. Moreover, problems may arise when there is a lack of available real-world data for training. Such data may be too expensive to collect or there may be privacy issues around data collection, and so simulated data are used instead to train the model. However, simulation, even with the introduction of stochastic processes to mimic naturally occurring variation, is only an approximation to the real world, and the application of the resulting model may produce unintended consequences.

Another characteristic of real-world data is that it is subject to temporal changes. ML models can be periodically retrained on new or revised data, which results in a change in the predictive nature of the model. Repeat certification may be quicker than the initial certification process, as many of the audit points around the qualitative review of the ML system (for example detailing information security and quality management processes) will remain unchanged. Whilst some larger ML models (including foundation models and LLMs such as ChatGPT) have a release cycle of months to a year, other applications can be retrained on a much more frequent basis. Existing cloud computing platforms such as Amazon AWS [30] and Google Cloud Platform [31] already allow for daily retraining of ML models within their off-the-shelf ML workflow tools, in a process known as continuous training and deployment. Consideration is required on whether continuous certification would be possible to match the fast-paced development of AI tools. This would require automated assessment methods for trustworthy AI metrics to be embedded within the software deployment cycle, much in the same way standard software performance tests are integrated into continuous deployment pipelines prior to pushing the code to a production environment [32]. From an end-user perspective the appropriate software auditing tools would need to be available to be able to determine what version of the model was in effect at any given point in time. Without this, transparency would be lost, as to why the same prompts might produce different outcomes on different days.

Certification of AI systems should include, at a minimum, certification of the trustworthy AI characteristics. An agreement on the metrics and definitions is yet to be reached [5]. Considering robustness as a measure of trustworthy AI, it may be possible to obtain statistical bounds for the robustness of a machine learning algorithm, such as a NN, through

empirical simulation on the increasing levels of noise within the model inputs [7]. However, a similar statistical approach to quantify the effects of deliberate, adversarial model attacks may not be as straightforward, especially with NN architectures that tend to learn the surface statistical regularities of the training data rather than higher level abstract representations [33].

Other developments in creating measures to quantify trustworthiness are still very much within the pre-normative stage; developing a comprehensive range of measures that are applicable to a wide variety of AI use cases may prove to be a significant measurement challenge. The US National Institute of Standards and Technology (NIST) have developed a risk management framework for AI to help address this challenge [4], following a direction from Congress. The guidance document is for voluntary use by organisations that are using AI across the whole deployment lifecycle, to integrate responsible practice into AI development and deployment with actionable guidance to operationalise trustworthy AI.

Finally it is also paramount that the human factors, being an integral part of the socio-technical characteristics of trustworthy AI, are taken into account, as the interaction between the AI system and humans using the system also demands testing and evaluation with measures beyond the traditional ones [34]. Human-AI interaction is inherently dynamic, with users learning the best ways to interact with an AI tool over time. And if the users are contributing to the training data set, the AI system will adapt as it gathers more interaction data. In addition, human usage of AI tools is naturally highly varied; it may not be possible to optimise a single model for all behaviour patterns, especially in cases where those behaviours are contradictory in nature. Novel evaluation frameworks are therefore required to assess human behaviour patterns with AI tools [35], and establish what affects these trends have on the trustworthy characteristics described above.

4 AI STANDARDISATION AND CERTIFICATION IN THE CONTEXT OF TRUSTWORTHY AI

Harmonisation of standards development between international, European, and national standards organisations is key for the implementation of safe and trustworthy AI systems. Here we present a brief overview of the latest development of documentary standards for AI from relevant standards bodies.

The EU AI Act (AIA 2021)

This legislative framework defines a set of objective-based requirements that an AI system must comply with, and provides requirements that scale with the level of risk in order to balance the rights and safety of consumers with the need for innovation [36]. The requirements are divided into the following areas:

1. Data and data governance: training, validation and testing data sets must meet a set of quality criteria. In particular, training, validation and testing data sets must have appropriate statistical qualities, be free of errors, and be representative of the intended use case.
2. Technical documentation: to be drawn up prior to market and kept up to date.
3. Record-keeping: automatic event logs, conforming to recognised standards or common specifications.
4. Transparency and provision of information to users: users can interpret the system’s output and use it appropriately.
5. Human oversight: the inclusion of human-machine interface tools.
6. Accuracy, robustness and cybersecurity: models to achieve an appropriate level of accuracy, robustness and cybersecurity in the light of their intended purpose, and to perform consistently with regards to these respects throughout their lifecycle.

7. Risk management system: effective and properly documented compliance with all regulatory requirements prior to product release, with robust quality and risk management systems and post-market monitoring.
8. Quality management system: to ensure regulatory compliance, through the accomplishment of the required conformity assessment procedure (certification).

The AIA categorises high-risk applications to be those primarily interacting with humans that have the largest potential to cause significant harm, providing a list of eleven identified operational areas in which certification and/or regulation will be required. Whilst the application area list may seem fairly comprehensive, there is a risk of the creation of regulation and standard gaps for novel use cases that do not fit into the predetermined application categories.

ISO/IEC JTC 1/SC 42

This ISO and IEC joint committee, founded in 2018, is responsible for standards development across the whole AI ecosystem, with a current total of fourteen published standards and an additional twenty-nine under development, as of March 2023 [37]. The deliverables cover a range of topics, from foundational concepts [38], to data quality and governance [39], reference architectures [40], use cases and application guidelines [41]. As the committee with the largest international membership it has the most comprehensive set of working groups and standards under production, though many of these are still within an early stage of development and it may be some time before these standards could be operationalised for conformity assessment.

IEEE 7000 series

The Institute of Electrical and Electronics Engineers (IEEE) have instigated a series of standards that have a strong focus on ethical aspects. The IEEE 7000 series of guides [42] focus on the areas of bias, transparency and accountability, and provide step-by-step instructions on how to adopt stakeholder values from the early development phases of an AI build through to deployment. The standards aim to provide ethical specifications which integrate seamlessly with system functional requirements.

ETSI

The European Telecommunications Standards Institute (ETSI) is another key player within the AI standards space, that has a focus on supporting not just the development of standards, but standards testing as well. Since the latter aspect requires a high degree of specification, the majority of ETSI's work is constrained to the application area of e-health [43]. ETSI also manage an Industry Specification Group (ISG) on Securing Artificial Intelligence (SAI) [44]; the committee is focussed on using AI to enhance security, mitigation against attacks that leverage AI, and securing AI systems against adversarial attacks.

ITU-T

The International Telecommunication Union Telecommunication Standardisation Sector (ITU-T) coordinates standards for information security. Like the ETSI standards, the ITU-T tend to be domain specific and narrow in scope, for example ITU Y.3172 [45] which describes an architectural framework to accommodate ML within 5G networks.

Overall, the ISO/IEC standards provide the most coverage of AI governance and applications, however in general AI standards are still very much at an early stage of development, and therefore most likely lagging with respect to development in AI applications and their real-world usage, as is often the case with fast-moving and disruptive

technologies. Finding the balance between the need for certification, on the one hand, and the need for innovation, on the other hand, is a dilemma that governments need to address in deciding what policies they will promote. In [46] a comprehensive review of AI standards development was performed through the lens of the EU AI Act requirements, using text mining techniques to determine the degree of alignment to the AIA. The authors also calculated an operationalisation index to provide a quantitative estimation of how relevant a standard is in turning the abstract AIA requirements into observable rules and features that could be used for certification. The ISO/IEC standards were found to have the highest levels of operationalisation, particularly in the areas of AI system life cycle management, functional safety, ML classification performance and the assessment of the robustness of NNs. On balance, the analysis showed that the existing standards (and those in development) provide comprehensive cover for requirements 4-8 in the list in the section above, but that there are significant gaps for the first three areas which primarily deal with data and data governance.

5 CHATGPT AND TRUSTWORTHY AI

ChatGPT was released in November 2022 by OpenAI, a research and deployment company [47]. The acronym GPT stands for generative pre-trained transformer; generative implies that the NN is able to generate content it has not previously seen, pre-training implies that although the model was trained on a next word prediction task it is intended for use on a variety of other tasks, and transformer refers to the NN architecture used to train the model. The GPT part is a pre-trained LLM with 175 billion parameters trained on a very large web crawl data set, including Wikipedia and other text corpora. The conversational (i.e. chat) part of ChatGPT was trained on top of GPT, using reinforcement learning, from human labelled demonstration data created by OpenAI.

It has several limitations, which OpenAI is gradually addressing:

1. It may give incorrect answers, one reason being that the data it is trained on did not have knowledge of the question.
2. It may be sensitive to small changes in the input question, so may need rephrasing to correctly answer the question.
3. The model may be biased for certain questions.
4. It may provide inappropriate answers.

Despite these limitations it is estimated that as of January 2023 ChatGPT reached the astonishing number of 100 million active users [48]. To name a few tasks, ChatGPT can write essays, help debug code, and answer open questions on a very wide variety of topics [49]. It is heralded as a potentially disruptive technology that could, for example, affect the way search engines interact with users, and more generally revolutionise automated question answering platforms.

LLMs such as ChatGPT are foundation models [28] which are a more general class of ML models that are trained on a vast amount of unlabelled multimodal data such as text, image, video and audio. A foundation model can be adapted by a mechanism known as fine-tuning to a wide range of downstream tasks, which are the tasks at hand to be solved using the model. Foundation models are thus empowered by transfer learning and scale. Transfer learning enables knowledge learnt from one domain to be used in another, and scale gives the model breadth of knowledge in a diverse range of domains.

There are several issues regarding ChatGPT and other foundation models that should be debated in the research community [50]. One issue raised is that of inaccuracy and bias in some results returned by ChatGPT. The authors propose that there is still a need for human verification of results, when used by researchers. Another issue is to do with accountability and transparency in the sense that authors using ChatGPT should be clear in their writings

when it was used and what for. In addition, a call is issued to invest in open LLMs, as ChatGPT and other LLMs are commercial products and thus lack full transparency. One such example is BLOOM [51] which is a multilingual open source LLM having 176 billion parameters. Nevertheless, the barrier to developing competitive open foundation models is the vast resource needed to construct, maintain and operationalise high quality foundation models.

Conversational AI models such as ChatGPT can also be used to make complex concepts accessible and understandable to non-experts, as suggested for the financial domain [52] but such document/concept summarisation techniques are applicable to other domains as well. Combining explainable AI methods with foundation models such as ChatGPT can provide better explanation than purely relying on ChatGPTs raw responses and thus increase users' trust in the AI system.

There are several examples where trustworthy AI principles are potentially violated in ChatGPT [53]. Regarding privacy, it is possible that training data with personal information may be extracted, by using the output of a ChatGPT prompt within an internet search. This is known as an inversion attack. There is also a growing backlash from writers (and artists regarding image generation models) around potential copyright infringement. ChatGPT can return text without crediting a source, copied from authors who placed their work on the internet in previous times with no concept of how AI might use it in the future. Regarding bias, it is argued that, for example, gender and racial bias may be present in the language model, since they are inherently present within the text corpora on which ChatGPT was trained. There is also the potential for inappropriate language to be present in the output, originating from trained text, although some LLM providers are addressing this with automated content moderation at both the model input (user prompt) and output (LLM response) points [54].

These issues mean that ChatGPT may violate fairness. The principle of fairness is to ensure that all social groups are treated equally without any discrimination. Fairness is a wider characteristic than bias-freeness in that it stipulates that AI systems should not cause harm to society and more generally should be accessible to all parts of society without discrimination. This is especially relevant to ChatGPT and other foundation models which may be used by a large and diverse group of users. Another major issue for foundation models is who may be accountable when things go wrong. As a prerequisite for ensuring accountability, it is necessary to have a traceable audit chain from data collection, through ML algorithm design, model construction and evaluation, and finally deployment.

We will now discuss a suggestion of how to regulate foundation models in the context of trustworthy AI, concentrating on ChatGPT as an operational LLM [55]. Direct regulation of AI systems, as proposed by the EU at the time of writing, suggests that an AI system used in a high-risk application will accordingly be subjected to high-risk obligations. It therefore follows that "general-purpose AI systems" (GPAIS) such as LLMs will be subjected to high-risk obligations if they are used in a high-risk application. However, the question remains to identify accountability for any problems that occur. There are at least four actors involved in the AI system value chain: (i) developers who create LLMs, (ii) deployers, who are fine-tuning LLMs for their specific use case, (iii) users who are generating outputs from LLMs, and (iv) recipients who consume a product generated by user(s).

From the developer point of view, due to the model complexity, it would be impractical to list all the risks the model entails within every possible specific usage domain, for example health care. Similarly, ensuring privacy and security in the use of all such derivative systems would not be feasible. This would put an undue burden of compliance to direct regulation of GPAISs on the providers of LLMs. It is therefore argued that regulation should concentrate on model deployers and users of LLM technology rather than on the LLMs developers

themselves. In particular, regulation needs to address the capability of LLMs to generate inappropriate and/or unsafe content including “fake news”. The democratisation of AI tools enables step changes in productivity not just for legitimate users, but also bad faith actors, in ways that will be difficult to legislate against.

It is worth mentioning that existing laws including non-discrimination and data protection, will still apply to LLMs within the AI value chain. There is also the requirement for content moderation of LLM outputs to enable the detection and blocking of harmful content. Developers, deployers and users should be required to report incidents and mitigation strategies used to deal with harmful content, which may require the development of new procedures to integrate with existing regulation on harmful/illegal internet content. We stress that, in principle, developers are still responsible under existing regulation for the content used in their products, and therefore must mitigate against undesirable side effects that may occur due to such content. In summary a technology-neutral approach to regulation is better suited for LLMs, that concentrates on the outcomes that are present in high-risk applications, or those pertaining to non-discrimination or data protection issues. This is accentuated by the fact that AI technology is moving at such as a high pace that regulating the underlying technology as such would not be practical.

Since certification includes the process by which products are assessed as conforming to standards, all the above considerations will apply with regards to the application of any regulatory standards, concentrating on the ability of the LLM deployers and users to demonstrate compliance, rather than the LLM creators. The creation of methods for testing fairness, for example, in such systems will be a technical challenge, given that the end user can create any possible input into the model. It would be relatively easy to come up with a set of expected input prompts for the application in question (for example, a help desk chat model for a bank), and compare the model output for combinations of questions and protected characteristics as and when they occur within the input questions (such as the use of gendered pronouns within the questions). However, the way diversity characteristics may present within LLM prompts may be far more subtle than this, with end users that use slightly different choices in wording, grammar and syntax, whilst still asking the fundamentally the same question of the model as was constructed by the test engineer. Regardless, to enable the ability for LLM product certification in high-risk areas, it will be necessary to address the formation of such testing and verification procedures.

6 CONCLUSIONS

The AI standards landscape is still evolving and in flux, playing catch-up with the speed at which AI and ML technology is progressing. Several key standards, such as ISO/IEC DIS 42001 Artificial Intelligence Management System [1], are already in draft form or just published, and there is a need for third-party certification to support the responsible use of AI in organisations with respect to their AI systems. Certification of AI systems is a key assurance service in building trust in AI and ML as laid out in the CDEI AI assurance roadmap [56]. Thus, concrete certification schemes need to be developed to ensure that AI systems are trustworthy.

We argued that certification should assess conformity of an AI system along three dimensions: (i) against standards such as 42001, (ii) against criteria pertaining to the trustworthiness of the system, and/or (iii) conformity testing and evaluation of the AI components of the system.

To substantiate our reasoning we have reviewed trustworthy AI and ML in its capacity to provide the key principles for responsible development and deployment of AI systems, and we summarised the landscape of AI standards and certification. As a case study we have discussed ChatGPT, a large language model that is generating a lot of attention and is

posing some difficult questions with respect to potential assessment conformity to the three pillars of trustworthy AI.

We have also briefly reviewed research on trustworthy AI and ML at NPL, emphasising the importance of uncertainty quantification as an essential component in delivering trustworthy AI, as it enhances the transparency of an AI system. Moreover, the strategic research agenda of the European Metrology Network for Mathematics and Statistics in Metrology, which NPL is a part of, emphasises the importance of a quality framework for AI systems and the specification of a standard interface for benchmarking, validation and certification of ML models.

7 REFERENCES

[1] International Organization for Standardization. (Draft 2023) Information technology — Artificial intelligence — Management system (SO/IEC DIS 4200) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:42001:dis:ed-1:v1:en

[2] International Organization for Standardization. (2022) Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (ISO/IEC 23053:2022) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:23053:ed-1:v1:en

[3] ChatGPT from Open AI https://chat.openai.com/chat (Accessed March 2023)

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

[5] Agarwal, A., Agarwal, H. & Agarwal, N., “Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems.” AI Ethics 3, 267–279 (2023). https://doi.org/10.1007/s43681-022-00147-7

[6] Irfan, M. M., Ali, S., Yaqoob, I. and Zafar, N., "Towards Deep Learning: A Review On Adversarial Attacks," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 91-96, doi: 10.1109/ICAI52203.2021.9445247

[7] Anderson, B. G., Gautam, T. and Sojoudi, S., “An overview and prospective outlook on robust training and certification of machine learning models.” IFAC Symposium on System Structure and Control (SSSC), 2022, doi: https://doi.org/10.48550/arXiv.2208.07464

[8] International Organization for Standardization. (2023) Information technology — Artificial intelligence — Guidance on risk management (ISO/IEC 23894:2023) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:23894:ed-1:v1:en

[9] Winter, P. M., et al., “Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications.” ArXiv abs/2103.16910 (2021)

[10] Ghosh, S. S. et al., “Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI.” ArXiv abs/2106.01410 (2021)

[11] JCGM GUM-6:2020 Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models

[12] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q., "On Calibration of Modern Neural Networks," Proceedings of the International Conference on Machine Learning, p. 1321–1330, 2017.

[13] Bhatt, U. et al., “Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty.” In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). Association for Computing Machinery, New York, NY, USA, 401–413 (2021). https://doi.org/10.1145/3461702.3462571

[14] Thompson, A. et al., “Uncertainty Evaluation for Machine Learning”, NPL Report MS 34, 2021, doi: https://doi.org/10.47120/npl.MS34

[15] Kendall, A. and Yarin, G., “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” Advances in Neural Information Processing Systems, vol 30 (2017)

[16] Brown, R. J. C., Janssen, J.-T. and Wright, L., “Why a digital framework for the SI?”, Measurement, 187, p110309 (2022) https://doi.org/10.1016/j.measurement.2021.110309

[17] Brown, C. et al., "Infrastructure for Digital Calibration Certificates," 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, Roma, Italy, 2020, pp. 485-489, doi: 10.1109/MetroInd4.0IoT48571.2020.9138220.

[18] European Metrology Network for Mathematics and Statistics, Strategic Research Agenda https://www.euramet.org/european-metrology-networks/mathmet/strategy/strategic-research-agenda (Accessed March 2023)
[19] The UK's National Quality Infrastructure https://www.gov.uk/guidance/the-uks-national-quality-infrastructure (Accessed March 2023)
[20] Swann, G. (2010), "International Standards and Trade: A Review of the Empirical Literature", OECD Trade Policy Papers, No. 97, OECD Publishing, Paris, https://doi.org/10.1787/5kmdbg9xktwg-en
[21] Cihon, P. et al., "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries," in IEEE Transactions on Technology and Society, vol. 2, no. 4, pp. 200-209, Dec. 2021, doi: 10.1109/TTS.2021.3077595
[22] Heesen, J., Müller-Quade, J., Wrobel, S. et al. "Certification of AI systems – Compass for the development and application of trusted AI systems" White Paper, Lernende Systeme – Germany's Platform for Artificial Intelligence, 2020 https://www.plattform-lernende-systeme.de/publikationen.html
[23] International Organization for Standardization. (2015) Quality Management Systems - Requirements (ISO standard no. 9001:2015) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-5:v1:en
[24] International Organization for Standardization. (2022) Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC standard no. 27001:2022) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:27001:ed-3:v1:en
[25] Galan, C., "The Certification as a Mechanism for Control of Artificial Intelligence in Europe" (September 11, 2019). http://dx.doi.org/10.2139/ssrn.3451741
[26] Balahur, A. et al., "Data quality requirements for inclusive, non-biased and trustworthy AI. Putting-Science-Into-Standards", Publications Office of the European Union, Luxembourg, 2022, doi:10.2760/365479, JRC131097
[27] Matus, K. J. M. and Veale, M. "Certification systems for machine learning: Lessons from sustainability" Regulation & Governance, 16(1), 177-196 (2021) https://doi.org/10.1111/rego.12417
[28] Bommasani, R. et al., "On the Opportunities and Risks of Foundation Models." ArXiv abs/2108.07258 (2021)
[29] Wing, J., "Trustworthy AI", Communications of the ACM, 64(10), pp 64-71 (2021) https://doi.org/10.1145/3448248
[30] AWS Machine Learning Guide https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html (Accessed March 2023)
[31] Google Cloud Platform Introduction to Vertex AI https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform (Accessed March 2023)
[32] Shahin, M., Ali Babar, M. and Zhu, L., "Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices," in IEEE Access, vol. 5, pp. 3909-3943, 2017, doi: 10.1109/ACCESS.2017.2685629.
[33] Jo, J. and Bengio, Y., "Measuring the tendency of CNNs to Learn Surface Statistical Regularities." ArXiv abs/1711.11561 (2017)
[34] Freeman, L. et al., "Best Practices for Addressing New Challenges in Testing and Evaluating Artificial Intelligence Enabled Systems", AIRC Perspectives, September 2022
[35] Chen, Q. Z., et al., "HINT: Integration Testing for AI-based features with Humans in the Loop. In 27th International Conference on Intelligent User Interfaces (IUI '22)". Association for Computing Machinery, New York, NY, USA, 549–565 2022) https://doi.org/10.1145/3490099.3511141
[36] The Artificial Intelligence Act https://artificialintelligenceact.eu/ (Accessed March 2023)
[37] ISO/IEC JTC 1/SC 42 Artificial Intelligence https://jtc1info.org/sd-2-history/jtc1-subcommittees/sc-42/ (Accessed March 2023)

[38] International Organization for Standardization. (2022) Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (ISO/IEC 22989:2022) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en
[39] International Organization for Standardization. (2022) Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations (ISO/IEC 38507:2022) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:38507:ed-1:v1:en
[40] International Organization for Standardization. (2020) Information technology — Big data reference architecture (SO/IEC TR 20547-1:2020) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-1:ed-1:v1:en
[41] International Organization for Standardization. (2021) Information technology — Artificial intelligence (AI) — Use cases (ISO/IEC TR 24030:2021) Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24030:ed-1:v1:en
[42] Institute of Electronic & Electrical Engineers (2021) IEEE Standards Model Process for Addressing Ethical Concerns during System Design (IEEE 7000-2021) Retrieved from https://standards.ieee.org/ieee/7000/6781/
[43] ETSI eHEALTH https://www.etsi.org/technologies/ehealth (Accessed March 2023)
[44] Industry Specification Group (ISG) Security Artificial Intelligence (SAI) https://www.etsi.org/committee/sai (Accessed March 2023)
[45] International Telecommnication Union. (2019) Architectural framework for machine learning in future networks including IMT-2020 (ITU-T Y.2172(06/2019)) Retrieved from https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=13894
[46] Nativi, S., and De Nigris, S., “AI Standardisation Landscape:state of play and link to the EC proposal for an AI regulatory framework”, EUR 30772EN, Publications Office of the European Union, Luxembourg, 2021,ISBN 978-92-76-40325-8, doi:10.2760/376602, JRC125952
[47] Introducing ChatGPT https://openai.com/blog/chatgpt (Accessed March 2023)
[48] Reuters Analyst Note: ChatGPT sets record for fastest-growing user base https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Accessed March 2023)
[49] Taecharungroj, V. “What Can ChatGPT Do?” Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data Cogn. Comput. 2023, 7, 35. https://doi.org/10.3390/bdcc7010035
[50] van Dis, E. A. M. et al., “ChatGPT: five priorities for research”, Nature 614, 224-226 (2023) doi: https://doi.org/10.1038/d41586-023-00288-7
[51] Introducing The World’s Largest Open Multilingual Language Model: BLOOM https://bigscience.huggingface.co/blog/bloom (Accessed March 2023)
[52] Yue, T., Au, D., Au, C. C. and Iu, K. Y., “Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology” (February 2023) http://dx.doi.org/10.2139/ssrn.4346152
[53] Li, P. “Trustworthy Natural Language Processing” http://lipiji.com/slides/TrustNLP.pdf (Accessed March 2023)
[54] Microsoft claims its new tools make language models safer to use https://techcrunch.com/2022/05/23/microsoft-claims-its-new-projects-make-language-models-safer-to-use (Accessed March 2023)
[55] Hacker P., Engel A. and Mauer M,, “Regulating ChatGPT and other Large Generative AI Models.” ArXiv.abs/2302.02337 (2023)
[56] CDEI “The roadmap to an effective AI assurance ecosystem” https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem (Accessed March 2023)