OPEN ACCESS
IOP Publishing | ISECAEP
Journal of Reliability Science and Engineering
J. Reliab. Sci. Eng. **1** (2025) 022001 (24pp)
https://doi.org/10.1088/3050-2454/adcef2

## Topical Review

# Reliability engineering, risk management, and trustworthiness assurance for AI systems

**Xiaoge Zhang**[1,*], **Tao Wang**[1], **Lei Ma**[2,3] **and Sankaran Mahadevan**[4]

[1] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong Special Administrative Region of China, People's Republic of China

[2] The University of Tokyo, Tokyo 113-8658, Japan

[3] The University of Alberta, Edmonton, AB, T6G 1H9, Canada

[4] Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN 37235, United States of America

E-mail: xiaoge.zhang@polyu.edu.hk, seatao.wang@connect.polyu.hk, ma.lei@acm.org and sankaran.mahadevan@vanderbilt.edu

Received 13 January 2025, revised 16 April 2025
Accepted for publication 21 April 2025
Published 29 April 2025

**Abstract**

As the potential applications of artificial intelligence (AI) continue to expand, a central question remains unresolved: will users trust and adopt AI-powered technologies? Since AI's promise closely hinges on the perceptions of its trustworthiness, how to guarantee the reliability and trustworthiness of AI plays a fundamental role in fostering its broad adoptions in practice. However, the theories, mathematical models, and methods in reliability engineering and risk management have not kept pace with the rapid technological progress in AI. As a result, the lack of essential components (e.g. reliability, trustworthiness) in the resultant models has emerged as a major roadblock to regulatory approval and widespread adoptions of AI-powered solutions in high-stakes decision environments, such as healthcare, aviation, finance, nuclear power plant, to name a few. To fully harness AI's power for automating decision making in these safety-critical applications, it is essential to manage expectations for what AI can realistically deliver to build appropriate levels of trust. In this paper, we focus on functional reliability of AI systems in the regime of supervised learning and discuss the unique characteristics of AI systems that necessitate the development of specialized reliability engineering and risk management theories and methods to create functionally reliable AI systems. Next, we thoroughly review five prevalent engineering mechanisms in the existing literature for approaching functionally reliable and trustworthy AI, including uncertainty quantification (UQ) composed of model-based UQ and model-agnostic conformal prediction, failure prediction, learning with abstention, formal verification, and knowledge-enabled AI. Furthermore, we outline several research challenges and opportunities related to the development of reliability engineering and trustworthiness

* Author to whom any correspondence should be addressed.

assurance methods for AI systems. Our research aims to deepen the understanding of reliability and trustworthiness issues associated with AI systems, and spark researchers in the field of risk and reliability engineering and beyond to contribute to this area of study with emerging importance.

Keywords: deep learning, reliability engineering, risk management, trustworthy AI, AI systems

(Some figures may appear in colour only in the online journal)

## 1. Introduction

Artificial intelligence (AI), particularly deep learning, has achieved extraordinary successes in profoundly transforming a wide spectrum of fields ranging from machine translation and object recognition to cancer diagnosis and prognostics as well as health management [1–3]. In essence, the enormous benefits enabled by AI are attributed to the powerful capability of neural networks in automatically discovering representations and patterns needed for detection or classification tasks by learning from a large volume of natural data in its raw form [4]. In traditional machine learning (e.g. support vector machine, decision tree), careful engineering and considerable domain expertise is required to engineer handcrafted feature extractor to convert raw data into appropriate feature representations to be consumed by the machine learning model. The emergence of deep learning has profoundly altered this situation in that it frees users and developers from tedious and challenging feature engineering task by eliminating the need of explicitly programming the rules for feature engineering. The end-to-end representation learning ability empowered by deep neural network largely lowers the bar of applying machine learning, reduces the effort of implementing data-driven solutions and thus makes deep neural network a top pick in the context of learning problems across a wide range of contexts.

While the end-to-end representation learning of deep neural network undoubtedly gives rise to breakthroughs in tackling a series of long-standing problems that are difficult to be formally defined in mathematical terms (e.g. image classification, machine translation, video analytics), the hierarchical nature of automatic representation learning in neural network provides limited interface for human to participate and intervene. Furthermore, these learned representations are hard to be verified and analyzed. These shortcomings result in a series of problems in the downstream decision-making activities, such as interpreting the reasoning mechanism, assessing the reliability of AI model, analyzing the behavior of AI models, understanding the failure modes of AI model, and evaluating the adverse outcomes of AI models. These concerns might not that matter in risk-free decision environments, such as machine translation, face recognition, while the situation is entirely different in safety-critical applications (e.g. power grid, aviation, self-driving). A common characteristic shared by high-stakes applications is their stringent operational conditions and strict safety requirements with an extremely low tolerance of errors caused by AI. In safety-critical applications like autonomous driving, a confidently wrong decision from deep learning model can violate safety and reliability standards, leading to catastrophic consequences and even fatal accidents [5, 6]. In particular, the high-stakes applications tend to prioritize safety over efficiency as the overriding criterion to consider [7–9]. The lack of mature solutions to assure the reliability and safety of AI-based solutions has significantly hindered the potential of AI in these critical decision-making environments [10–12]. As such, it renders a strikingly low rate of translating AI models into practical solutions in high-impact decision settings [13]. Take healthcare as an example. Only a limited number of AI-based solutions have been approved by pertinent regulatory agencies for use without human oversight, and most of the approved applications are restricted to low-risk settings [14, 15].

On the surface, the inherent deficiencies in deep learning escalate and manifest themselves as safety, reliability, and trustworthiness-related issues. In fact, several recent studies have made attempts to surface the limitations of deep learning models in various contexts [16]. For example, Nguyen *et al* [17] found out that deep neural networks with state-of-the-art performance could be easily fooled in that they classified many images that were completely unrecognizable to humans with over 99% confidence as members of a recognizable class (e.g. labeling with certainty that TV static is a motorcycle). Majumder *et al* [18] revealed that production-grade generative pre-trained transformer (GPT) model failed to comply with strict safety performance guidelines in power grid (e.g. voltage magnitude limits) and compromised the safe operations of electrical grid if GPT was integrated into power systems. Hager *et al* [19] found out that state-of-the-art large language models (LLMs) exhibited poor adherence to well-established diagnostic and treatment guidelines and were sensitive to changes in prompts as well as the order of information presented to the model, thus posing a serious risk to the health of patients. Besides, LLMs are also criticized for generating overly confident yet factually nonsensical content known as hallucinations in the literature [20, 21]. The hallucinations have resulted in several failure cases of generative AI systems as reported in the literature [22]. Last but not least, as deep learning is built upon the independent and identically distributed (i.i.d.) foundation [23], the resultant AI model is limited in its capability to handle inputs from a distribution different than the training data. If an AI model is used to process novel situations or corner cases it never encounters before (e.g. out-of-distribution, OOD), the model is prone to generating erroneous predictions because the inputs go beyond the scope of the trained model. For example, a driver-less ride-hailing

car from Baidu hit a pedestrian in Wuhan reportedly crossing the street against the traffic light in China as the autonomous driving software failed to handle this unconventional behavior breaking traffic laws [24]. This accident clearly highlighted the vulnerability and limitations of AI-based autonomous driving software in managing unconventional real-world situations, such as vehicles or pedestrians violating traffic laws.

The scenarios described above are just the tip of the iceberg of the pressing reliability and safety concerns surrounding AI. Since AI-enabled systems exhibit unique characteristics distinct from conventional engineering systems (e.g. civil, mechanical, electrical systems) that reliability modeling and theories have been traditionally developed for, including large number of model parameters, high-dimensional input space, no clear definition over the feasible region of model inputs, opaque reasoning mechanism, massive state space, existing traditional reliability modeling theories and methods, such as Bayesian network, fault tree, event tree, failure mode, effects & criticality analysis, state space model, physics of failure, are no longer applicable in the context of AI. Due to the exponential growth of the state space over neurons and activation functions in neural network, the traditional reliability engineering methods fundamentally lose their applicability due in part to poor scalability. Besides, the distinct characteristics of AI necessitate the development of reliability engineering theories and risk management methods designed specifically for AI. To ensure responsible use and safe utilization of AI in mission-critical applications, reliability engineering and risk management are crucial in understanding and managing over what AI can realistically achieve. Such studies are essential for unlocking AI's potential in an orderly, responsible, and controlled manner.

In this paper, we are motivated to give an overview on reliability engineering, risk management, and trustworthiness assurance for AI systems and suggest possible avenues ahead to move towards reliable, trustworthy, and controllable AI systems by equipping AI with layered protection against heterogeneous sources of risks. In our study, AI systems refer to the core AI model along with its desired operational setting, we focus on the functional reliability of its decision-making capability rather than encompassing hardware or the entire physical infrastructure it might be part of (like a robot). Towards this end, this paper investigates several engineering measures and strategies to ensure the functional reliability of AI systems developed under the supervised learning paradigm during normal use. Formally, functional reliability refers to an AI system's ability to accurately perform its intended functions with an acceptable level of reliability. Mathematically, a functionally reliable AI system can be defined as $\mathbb{P}(\mathcal{Q}(f(X), Y) \in \mathbb{S}) > \mathcal{R}$, where $f(X)$ represents AI system's output for the input $X$, $Y$ denotes the desired or intended output, $\mathcal{Q}$ is a task-specific function to evaluate the performance of AI system's output in the specified operation context, $\mathbb{S}$ is a set defining the range of acceptable or satisfactory performance level, and $\mathcal{R}$ denotes the reliability requirement for the fitted model $f(X)$ specified by end user (e.g. 0.99, 0.999). Note that $X$ can be randomly generated multiple times, we consider $f(X)$ to be functionally reliable only if the above reliability constraint is satisfied for all realizations of $X$ across all runs. Take an AI-enabled autonomous vehicle as an example. When the vehicle makes a left turn at an intersection, the AI-based driving system needs to determine the appropriate turning angle and turning speed. In this context, $X$ denotes the current position of the vehicle in the road and the sensed surrounding environment, including traffic signal, pedestrian, obstacles, surrounding vehicles, road conditions, etc. The function $f(X)$ denotes the action (e.g. turning angle, driving speed) taken by the AI system to execute the left turn while $Y$ represents the optimal or reference action that should be taken to perform the left turn safely and efficiently in ideal conditions. If safety, measured by the distance between obstacles and the vehicle, is our major concern in this context, then $\mathcal{Q}$ is a user-defined function to measure the minimum distance between surrounding objects and the vehicle, and $\mathbb{S}$ is a set of acceptable range for the margin of safety. Note that $\mathcal{Q}$ could also encompass other performance aspects relevant to making the left turn, such as the smoothness of the turn, adherence to traffic rules, efficiency (time taken for the turn), or passenger comfort. For the sake of simplicity, we suppose $\mathcal{Q}$ is one-dimensional function focused solely on measuring the safety margin. In this context, the AI-based driving system is considered reliable for executing the left turn if it meets the margin of safety requirement specified by $\mathbb{S}$ with a probability of at least $\mathcal{R}$.

Since this paper centers on the functional reliability of AI systems under normal use, research topics such as data privacy, ethical concerns, evasion, model poisoning, and adversarial attack are beyond the scope of this paper. In this paper, we review five major technical methods for approaching reliable and trustworthy AI, including uncertainty quantification (UQ), failure prediction, learning with abstention, formal verification, and knowledge-enabled AI. Note that our review of this topic and its subtopics is not exhaustive, as each subtopic is extensive enough to warrant its own independent review paper. The rest of this paper is organized as follows. In section 2, we define reliability engineering in the context of AI systems and elucidate several unique characteristics of risks associated with AI that cannot be identified and managed with the conventional reliability engineering and risk management methods. In section 3, we review several commonly adopted strategies in the extant literature for reliability engineering, trustworthiness assurance, and risk management of AI models in the open world. In section 4, we summarize the research challenges and opportunities in reliability engineering and trustworthiness assurance of AI. In section 5, we end this paper with concluding remarks and future research directions.

## 2. Reliability engineering, risk management and trustworthiness assurance for AI systems

According to Wikipedia, reliability is defined as 'the probability that a product, system, or service will perform its intended function adequately for a specified period of time, or

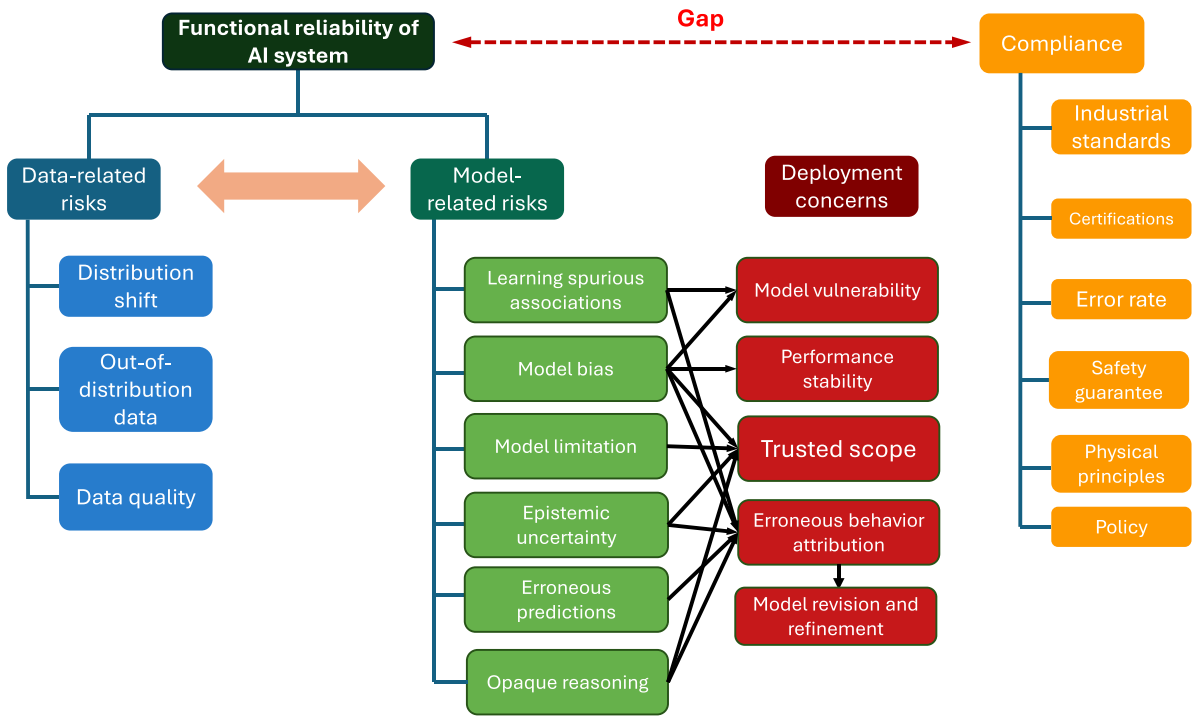


**Figure 1.** Overview of functional reliability of AI systems. At a high level, the functional reliability of AI systems faces two major sources of risks: data-related risk and model-related risks.

will operate in a defined environment without failure' [25, 26]. Analogously, in the context of AI systems, functional reliability is concerned with the overall response correctness of AI systems under the conditions of expected use and the assurance of AI system in performing its intended functions reliably. Since environment is an essential element in shaping the reliability of AI systems, reliability engineering of AI needs to be defined from two aspects: operation environment and model. Figure 1 provides an overview on the functional reliability of AI systems. At a high-level, the functional reliability of AI systems faces two major sources of risks: data-related risks (or environment) and model-related risks. Thus, it is important to ensure that AI is operated under a valid environment in the sense that the inputs to the model fall within the scope of the trained AI model. It is well-known that modern machine learning systems behave unreliably and underperforms when encountering novel data [27–30], and this poses a significant safety risk to ML-enabled services in high-stakes environments, such as healthcare, aviation, robots. On the other hand, even if the input data is within the scope of the trained AI model, AI model might still make mistakes in certain cases due to other reasons [31, 32], such as epistemic uncertainty arising from the lack of knowledge, learning of spurious correlations for reasoning, sample selection bias, etc. Thus, it is important to investigate how to understand the failure modes and pathological behavior of AI systems in producing erroneous predictions, and develop effective measures to detect prediction failures of AI systems in advance for preventing them from causing adverse outcomes [33]. Failure mode and effects analysis (FMEA) on AI will deepen our understanding on the limitations of the underlying system and their effects on the downstream engineering systems of interest, thus inspiring the development of reliability assurance strategies, such as fallback. As the focus of this paper is on reliability engineering and trustworthiness assurance of AI systems under normal use, issues on data privacy, ethical problems, and adversarial attack of AI are beyond the scope of this paper.

On the input data side, how to guarantee that the input fed into deep learning model falls within the scope of the trained model is a prerequisite to create functionally reliable AI systems. In particular, as deep learning is built upon the i.i.d. foundation [34], the resulting model is limited in its capability to handle inputs from a distribution dissimilar with the training data. The unknown situations arising from the model input pose substantial risks and might lead the model to generate misleading predictions. For example, if an image of bear is fed into a neural network trained for classifying dog vs cat, no matter the model labels the image as dog or cat, the conclusion is always wrong as the input goes beyond what the model is trained for. The aforementioned example is known as OOD in the context of deep learning [23, 35]. Since we have no control over what is fed into the trained model after its deployment, it is essential to develop safety guardrails to filter out input data breaking the i.i.d. condition. In addition to OOD, dataset shift between the model development and deployment environment, the difference in data acquisition device and techniques, population demographics, etc, also pose a threat to the functional reliability of AI systems [36]. Note that distributional shift is fundamentally different from OOD. Distributional shift occurs when

the statistical properties of the input data change between the training phase and the deployment phase. As a result, the data distribution the model encounters during deployment is different from the distribution the model is trained upon. In contrast, OOD data arises from a different domain or category that the model has not previously encountered. As the OOD data lies outside the support of the training data distribution, the model has not learned to recognize or process such data during training. Formally, distribution shift can be mathematically represented as $\mathbb{P}(\mathcal{X}_{\text{production}}, \mathcal{Y}_{\text{production}}) \neq \mathbb{P}(\mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}})$, while OOD data is modeled as $\mathbb{P}(\mathcal{Y}_{\text{train}} \cap \mathcal{Y}_{\text{OOD}}) = \emptyset$, where $\mathbb{P}(\mathcal{X}_{\text{production}}, \mathcal{Y}_{\text{production}})$ denotes the underlying distribution governing the observed production data after model deployment and $\mathcal{Y}_{\text{OOD}}$ denotes the distribution of target class for OOD data. Since the failures caused by OOD data and dataset shift are difficult to anticipate, a series of studies have witnessed a lot of progress on the detection of OOD and dataset shift at runtime [23, 37, 38]. Despite the rapid progress, there is still no mature solution to provide a formal theoretical guarantee on creating a safe and reliable environment for AI model to operate and deliver the promised outcome.

In addition to the risk posed by the model inputs, deep learning model also struggles with learning the right representations for inference, reasoning, and decision making. Several state-of-the-art studies have shown that deep learning tends to learn associations and biases rather than stable cause-effect relationships from the development data [39, 40]. Some associations learned by deep neural network are unstable and do not make any sense in reality. For example, Ribeiro *et al* [39] designed a dataset that all pictures of wolves had snow in the background, while pictures of huskies did not, used the dataset to train a logistic regression classifier, and found out that the resultant classifier predicted 'Wolf' if there was snow in the picture, and 'Husky' otherwise, regardless of the animal color, position, pose. Rather than learning the inherent difference between 'Wolf' and 'Husky', the trained AI model relied on spurious associations to arrive at decisions in the classification task. In this regard, large-language models (i.e. ChatGPT, Gemini) are no exception as they exhibit a tendency to generate overly confident yet factually nonsensical content, known as hallucinations in the literature [20, 21]. Such behavior arises from the learning of unreasonable associations and correlations from the raw data. These learned spurious associations and correlations inevitably increase the vulnerability of resultant AI models, and is detrimental to the reliability, stability, and generalization of AI systems in performance. In particular, the vulnerable learning paradigm could lead to catastrophic outcomes in high-stakes decision-making environments. Unfortunately, we still do not have well-established mechanisms to detect when the underlying AI model learns what kind of spurious correlations, understand the propagation of the learned spurious associations through the environment that AI is deployed in, and analyze the impact of these learned unstable relationships on the downstream decision-making activities.

The learned spurious relationships also increase the complexity of model performance diagnosis in neural network. That is, if AI models malfunction or generate erroneous predictions, how should we revise them such that they do not commit the same error repeatedly. This question is intricate in nature as we do not know how to attribute the erroneous behavior of neural network to the neurons and which of the billions of parameters need tweaking when the models make a mistake. In this regard, it is also important to highlight the difference in model update between traditional software systems and AI-enabled software/systems. In traditional software systems, as the logic of the software is explicitly programmed, developers can identify and fix the bug in the software with the help of troubleshooting tools, make corresponding changes in the source code, and observe how the updated software changes the performance of the software. In contrast, in the era of AI, these debugging tools fundamentally fail to work as they are incapable of debugging the behavior of AI models as there is no explicit rules and functions to define how AI behaves. In particular, as the behavior of AI model is a complex function of training data, model architecture, training method (e.g. optimization algorithm, hyperparameter tuning), and deployment environment, these elements, individually and collectively, play a crucial role in affecting the behavior of AI systems. In certain cases, although AI models can be retrained to fix their errors, retraining is unfortunately expensive and time-consuming, and adjusting all of the model's billions of parameters to fix a particular error would also be overkill. The opaque reasoning process of neural networks further complicates the troubleshooting of neural network model behavior. Specifically, there is still no well-established approach to edit neural network in a way that is not as computationally expensive as model retraining.

Even if we factor out the spurious associations and biases from raw data learned by the neural network, a well-trained AI model is not omnipotent as the model might still have high uncertainty about some edge in-distribution cases because it never encounters such cases in the training data. As it is impossible to build training data that encompasses all possible scenarios, the trained AI model will inevitably encounter edge cases after its deployment. For example, if we train a classifier for discriminating digits (e.g. 0, 1, 2, etc), the trained model might be uncertain if it is used to classify an ambiguous image with a number that is generated by mixing the images of 2 and 3 together. In such cases, to play safe, we hope that the model prediction is accompanied by a high epistemic uncertainty, where the estimated uncertainty serves as an effective communication channel to alert us about the unusual nature of the input data.

The preceding paragraphs elucidate the concept of reliability engineering, risk management and trustworthiness assurance for AI systems, and underscores the key difference in the reliability engineering between AI systems and other conventional engineering disciplines. As the potential applications of AI continue to expand in breadth and depth, how to manage the risks posed by AI itself as well as its integration into various safety-critical systems (e.g. power grid, autonomous driving, medical diagnosis) and ensure that AI delivers the outcome promised remain a central issue to be tackled. In particular,

how to manage the inherent risks in AI systems responsibly and ensure that AI complies with existing industrial standards and strict performance requirements of mission-critical applications with a correctness of response guarantee has become an imperative issue to be addressed. To harness the power of deep learning in these high-stakes decision settings, it is important to assure the reliability of the underlying system by managing for what AI can realistically deliver. To this end, traditional reliability engineering theory, modeling methods, fault diagnosis approaches, reliability optimization algorithms, such as Bayesian network, fault tree, event tree, FMEA, lag far behind and struggle to keep pace with the rapid development of AI. They are fundamentally inadequate for analyzing the complex behavior of neural networks, performing attribution, and updating AI systems due to the massive number of parameters and exponential growth of state space, complex structure of neural networks, and intricate interactions among neural network components. The unique characteristics of AI systems necessitate the development of scalable reliability engineering theories and risk management approaches specifically tailored to AI. Hence, this paper aims to draw the attention of reliability engineers and researchers to this critical research area, share some of our initial thoughts and ideas, and encourage contributions to the advancement of reliability engineering theories and practical methods in the context of AI, ultimately fostering the creation of safe, reliable, and trustworthy AI systems.

## 3. Potential avenues ahead

In this section, we review several prevalent solutions in the literature to construct guardrails for safeguarding AI's adoptions in high-stakes decision settings. These solutions represent the current state-of-the-art literature for approaching reliable and trustworthy AI systems. Figure 2 provides a high-level overview of existing methods for assuring the functional reliability of AI systems, including UQ, failure prediction, learning with abstention, formal verification of neural networks, and knowledge-enabled AI.

### *3.1. UQ*

As mentioned earlier, since AI is not omnipotent, it is not only important to let neural network learn what it knows, but also let neural network learn what it does not know. Rather than relying on misleading predictions produced by neural networks, we should let neural network abstain when it is not confident about its predictions. Neural networks developed under such learning paradigm are allowed to abstain whenever they are not sufficiently confident in their predictions. In this section, we review several state-of-the-art methods for adding a layered protection through UQ.

*3.1.1. Model-based UQ.* Prediction of neural network, without a measure of their veracity, do not provide the necessary trust needed to make decisions in critical applications. To build the trust, UQ extends the traditional discipline of statistical error analysis to capture uncertainties due to noisy data, missing and undetected dependencies, overlooked exogenous factors, model parameter uncertainty, and the discrepancy between model forms and modeling strategies [7]. By accounting for these diverse sources of uncertainties, UQ provides a unified approach to provide quantitative insights beyond what a deterministic model could offer. Essentially, the estimated uncertainty serves as a communication channel for the model to express what it does not know.

In the context of neural networks, we exploit UQ to explore and understand their limitations. By utilizing the quantified uncertainty, we determine when to trust the predictions made by neural network, when to hand the decisions over to domain experts for further examination. In UQ, we associate each deterministic prediction with an uncertainty estimation (e.g. standard deviation) via the development of probabilistic deep learning models. Unlike traditional neural network, probabilistic neural network is oftentimes modeled in a Bayesian fashion to quantify the parameter and structure uncertainty of neural network [41, 42]. In the Bayesian context, neural network parameters (e.g. weights, bias) are assumed to follow some predefined probability distributions (e.g. Gaussian distribution) rather than deterministic values, these assumed prior distributions are then combined with training data to infer posterior distributions of neural network parameters. After the full posterior distributions of neural network parameters are derived, the predictive distribution on an unseen data point can be obtained by integrating over the posterior distributions of neural network parameters.

Given the large number of parameters in the neural network, a variety of approximation methods, such as Laplace approximation [43], Markov chain Monte Carlo (MC) [44], variational Bayesian methods [45–47], have been developed to replace exact Bayesian inference for reducing the computational burden. Unfortunately, these methods require significant modifications to the training procedure of neural network and suffers from prohibitive computational cost and poor scalability. To tackle this computational challenge, a series of methods have been developed to make Bayesian inference scalable in neural network. For example, Gal and Ghahramani [48] developed MC dropout to approximate Bayesian inference in a particular setup, where the variational posterior of neural network weights was a Bernoulli mixture of two independent Gaussians of fixed covariance in MC dropout. Subsequently, they extended MC dropout to convolutional neural network and recurrent neural network for uncertainty estimation [49, 50]. Lakshminarayanan [51] addressed the issue of computational efficiency from the perspective of distributed computation by developing a simple and scalable approach for quantifying both aleatory and epistemic uncertainty in neural networks based on ensembles of neural networks. This method was the first to evaluate the quality of predictive uncertainty on the large-scale ImageNet dataset.

Recent efforts for estimating the predictive uncertainty of neural networks have shifted towards deterministic uncertainty estimation (DUE) methods. Distinct from MC dropout and deep ensemble, DUE is featured by its single forward pass

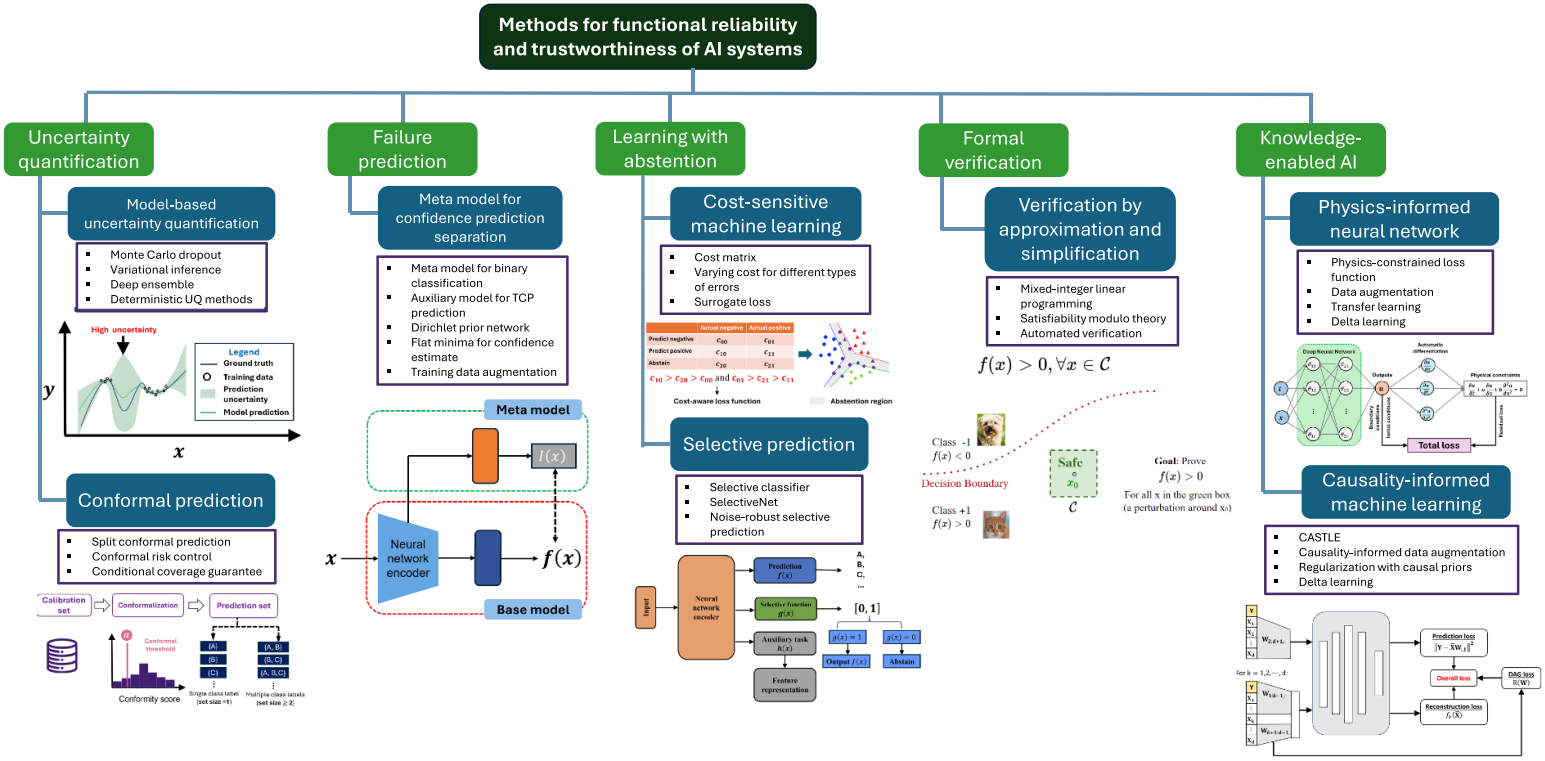


**Figure 2.** Structure of methodological review for assuring functional reliability of AI systems. The review paper covers five prevailing methods for approaching the functional reliability of AI systems: uncertainty quantification, failure prediction, learning with abstention, formal verification of neural networks, and knowledge-enabled AI.

characteristic for uncertainty estimation. Spectral-normalized Neural Gaussian Process (SNGP) [37], DUE [52] and deterministic UQ [53] are among the representative methods in this class of UQ methods. These DUE methods are featured by the seamless integration of distance-preserving feature extractor powered by neural network and efficient approximations to GP (e.g. random Fourier features-approximated GP, inducing points-based sparse GP) in an end-to-end trainable manner. In doing so, the resultant model exploits neural network's powerful representation learning capability to extend the application scope of GP to high-dimensional problems, while it also resembles GP's principled uncertainty estimation behavior in distance awareness.

Regardless of the specific UQ methodology or uncertainty measure (i.e. maximum class probability (MCP), entropy, standard deviation), the quantified uncertainty serves to differentiate between high-confidence and low-confidence model predictions after an uncertainty threshold is specified [23, 54, 55]. Low-confidence predictions are expected to be correlated with high prediction uncertainty and vice versa. Typically, the uncertainty estimate is expected to objectively convey the degree of confidence in the correctness of model predictions, and this is particularly important to high-stakes decision environments [7, 56, 57]. Established on this logic, UQ has been extensively utilized in the literature to detect distribution shift and OOD data. Reliable detection of instances breaking the i.i.d. condition, upon which machine learning is founded, is crucial for the safe and responsible deployment of AI in practice [58–61]. For example, Sensoy *et al* [62] tackled the uncertainty estimation problem of neural network from an evidence theory perspective, and designed a neural network to express its multinomial opinions on the classification of a given sample as a Dirichlet distribution, thereby equipping neural network with the ability to say 'I do not know.'; Ovadia *et al* [63] used a large-scale dataset to assess the quality of uncertainty estimates of several state-of-the-art UQ methods on classification problems under dataset shift. They found out that post-hoc calibration only gave good results in i.i.d. regimes, but failed under even a mild shift in the input data; Linmans *et al* [64] compared the uncertainty estimate of prevalent UQ methods on both in-distribution and realistic near and far OOD data on on large-scale digital pathology datasets and showed that uncertainty estimates could be used to discriminate in-distribution from OOD data with high area under the curve scores. Zhu *et al* [65] revealed that most confidence estimation methods were harmful for detecting misclassification errors and proposed to enlarge the confidence gap between in-distribution and OOD data by searching for flat minima. In doing so, the proposed optimization approach leveraged confidence-based uncertainty estimate to yield state-of-the-art failure prediction performance under a variety of settings including balanced, long-tailed, and covariate-shift classification scenarios. Corbière [66] proposed learning the true class probability (TCP) as an alternative to the MCP-based uncertainty measure to better represent model confidence in failure predictions. Recently, Mucsányi *et al* [67] presented the first systematic study on the disentanglement of uncertainty estimate and underscored the importance of developing task-centric and disentangled uncertainty estimators.

*3.1.2. Conformal prediction.* Conformal prediction is a distribution-free, model-diagnostic UQ approach that generates statistically valid prediction regions to support reliable prediction-powered inference with any machine learning model [68–70]. Conformal prediction typically functions as a separate post-hoc processing step to convert machine learning model predictions into valid prediction bands containing the true class with a user-specified coverage. In the literature, most conformal prediction methods adopt the split conformal prediction setting [71], where a held-out set is used to calibrate the model prediction and assess its level of confidence and limitations. Generally, we can categorize conformal prediction methods into three main streams.

**Conformal prediction under data exchangeability**: Conformal prediction is a framework for constructing prediction sets that provide finite-sample marginal coverage guarantees under the condition of data exchangeability [72]. Figure 4 illustrates the basic steps of conformal prediction in generating statistically valid prediction sets in the context of classification problems. Procedurally, let $\alpha \in (0,1)$ denote the rate that a given ML model is permitted to commit the error, the goal of conformal prediction is to create a prediction band $\mathcal{C}$ such that for any new exchangeable pair $(\boldsymbol{x}_{n+1}, y_{n+1}) \sim \mathcal{P}$, such that $P(y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{n+1})) \geqslant 1-\alpha$ always holds, where the probability is over all of our data $(\boldsymbol{x}_i, y_i)$, $i = 1, \cdots, n+1$ and $(\boldsymbol{x}_i, y_i)$, $i = 1, \cdots, n$ denote the data used to train the ML model. To establish this guarantee, we need to rely on a calibration set for defining a non-conformity score function $s(\boldsymbol{x}, y)$ to measure the strangeness of the data point. A larger $s(\boldsymbol{x}, y)$ indicates the data point $(\boldsymbol{x}, y)$ deviates from the trend that the model has learned from the training data [73]. For example, given a fitted model $f$, in the case of classification problems, a conformity score can be defined as $s(\boldsymbol{x}, y) = f(\boldsymbol{x})_y$, where $f(\boldsymbol{x})_y$ denotes the probability of the fitted model $f(\boldsymbol{x})$ in assigning the right class $y$ to the input $\boldsymbol{x}$; in the case of regression problems, residual score [74], defined as $s(\boldsymbol{x}, y) = |y - f(\boldsymbol{x})|$, is commonly used for quantifying the degree of nonconformity. A high residual value indicates that $(\boldsymbol{x}, y)$ does not align well with the behavior of the model trained upon the available data. Next, we calibrate the model prediction based on the non-conformity score $s(\boldsymbol{x}, y)$ and employ the $1-\alpha$ quantile of the nonconformity score over a calibration set as the threshold to create the prediction band $\mathcal{C}$. Figure 3 illustrates the basic steps of conformal prediction in converting point predictions produced by a polynomial model trained for a 1D regression problem into valid prediction intervals given a target coverage of 90%. Out of the 100 testing samples, the actual values of 92 samples fall within the prediction intervals constructed by conformal prediction. As a result, the prediction intervals constructed by conformal prediction are statistically valid as it successfully achieves the target coverage.

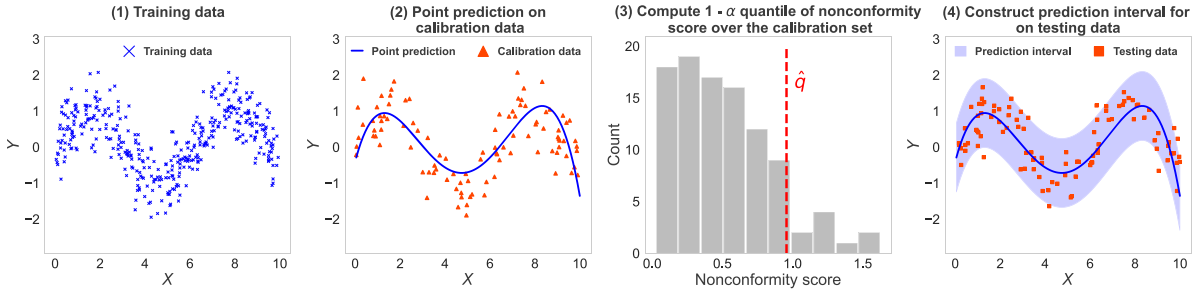


**Figure 3.** Demonstration of conformal prediction on top of a machine learning model fitted with polynomial regression over a 1D problem. (1) Training data: 400 samples are randomly generated from $Y = \sin(X) + \epsilon$, where $X \sim \mathcal{U}[0, 10]$ and $\epsilon \sim \mathcal{N}(0, 0.5^2)$. (2) Point prediction: Polynomial regression $f(x)$ with a degree of 4 is used to fit the 400 training points. The trained model is then used to make predictions on the 100 samples in the calibration set $\{(x_j, y_j)\}_{j=1}^{j=100}$. (3) Nonconformity score: we derive the nonconformity score $s(x, y) = |y - f(x)|$ for each sample in the calibration set. Given the miscoverage rate of $\alpha = 0.1$, we compute $1 - \alpha$ quantile of the nonconformity score over the calibration set denoted by $\hat{q}$. (4) Prediction interval construction: The point prediction $f(x)$ (blue line) for 100 randomly generated testing points are converted into valid prediction intervals (shaded light blue) $[f(x) - \widehat{q}, f(x) + \widehat{q}]$ by conformal prediction.

Formally, the performance of a conformal predictor is often evaluated using two indicators: statistical validity and the average set size (or width of prediction interval). Ideally, we prefer the set size (or width of prediction interval) of a conformal predictor to be as narrow as possible while achieving the target coverage. As conformal prediction guarantees the model coverage under data exchangeability, a lot of studies have concentrated on minimizing the average set size of conformal predictor. For example, Romano *et al* [75] introduced a conformality score to construct adaptive prediction sets for multi-class classification problems that enjoy provable finite-sample coverage by regularizing the non-conformity score to avoid unreliable probability in the tail. Later, Angelopoulos *et al* [76] proposed to regularize the small scores of unlikely classes after Platt scaling to generate smaller set size that contain the ground truth label with a formal finite-sample coverage guarantee.

**Conformal risk control under data exchangeability**: Conformal prediction has been extended to conformal risk control [77], where the loss is not limited to miscoverage but can incorporate any non-increasing, arbitrary loss function, such as error rate, false positive rate. In addition, Bates *et al* [78] developed PAC-conformal risk control to provide high-probability control with monotonic loss across any action space. Angelopoulos *et al* [79] further extended this approach to create machine learning models with finite-sample statistical guarantee for any non-monotonic loss function for any (unknown) data-generating distribution. All these methods operate under the data exchangeability assumption while extending conformal prediction to account for different types of loss beyond just coverage.

**Non-exchangeable, conditional coverage, and other extensions:** Recently, conformal prediction has been adapted to scenarios where the data exchangeability assumption does not hold [80]. This is particularly relevant for predictive models deployed in practical applications where data distributions vary over time. Established upon conformal risk control, Farinhas *et al* [81] adapted it for non-exchangeable data conditions. Some other notable variants focus on creating machine learning models satisfying conditional coverage. For example, Vovk [82] stated that it is impossible to achieve exact conditional coverage universally in finite samples. Therefore, most recent research concentrate on marginal coverage across a set of test points by relaxing the notions of conditional coverage. Notably, [83–85] explored the relationships between conformal prediction, group-wise coverage, and batch multi-valid coverage. For example, Jung *et al* [85] defined a spectrum of problems interpolating between marginal and conditional validity to address the gap between marginal and conditional coverage. Furthermore, Blot *et al* [86] extended conformal risk control to autonomously adapt models to the challenging test samples. Many other tasks in conformal prediction, such as class-condition methods [87], credal set prediction [88], and LLM factual guarantees [89], have also been actively investigated in recent years.

Table 1 compares the UQ methods reviewed earlier in this section from several dimensions, including the effort for recasting a deterministic neural network into the probabilistic counterpart, computational scalability in training uncertainty-aware models, computational efficiency for UQ at inference, statistical validity of uncertainty estimate, etc. From the quantitative comparisons in table 1, we make several interesting findings. First, traditional Bayesian inference methods, such as MCMC, variational inference, incur sophisticated modifications to the deterministic neural network and suffer from poor scalability when training uncertainty-aware models. Unlike the conventional Bayesian inference methods, MC dropout and deep ensemble are model-agnostic and provide significantly better scalability in training probabilistic models. However, the uncertainty estimate by MC dropout and deep ensemble is not statistically meaningful. Although calibration could enhance the statistical significance of the quantified uncertainty, it provides no guarantee on the statistical validity of the uncertainty estimate. Conformal prediction, on the other hand, holds significant potential for creating reliable and trustworthy AI, as it inherently provides

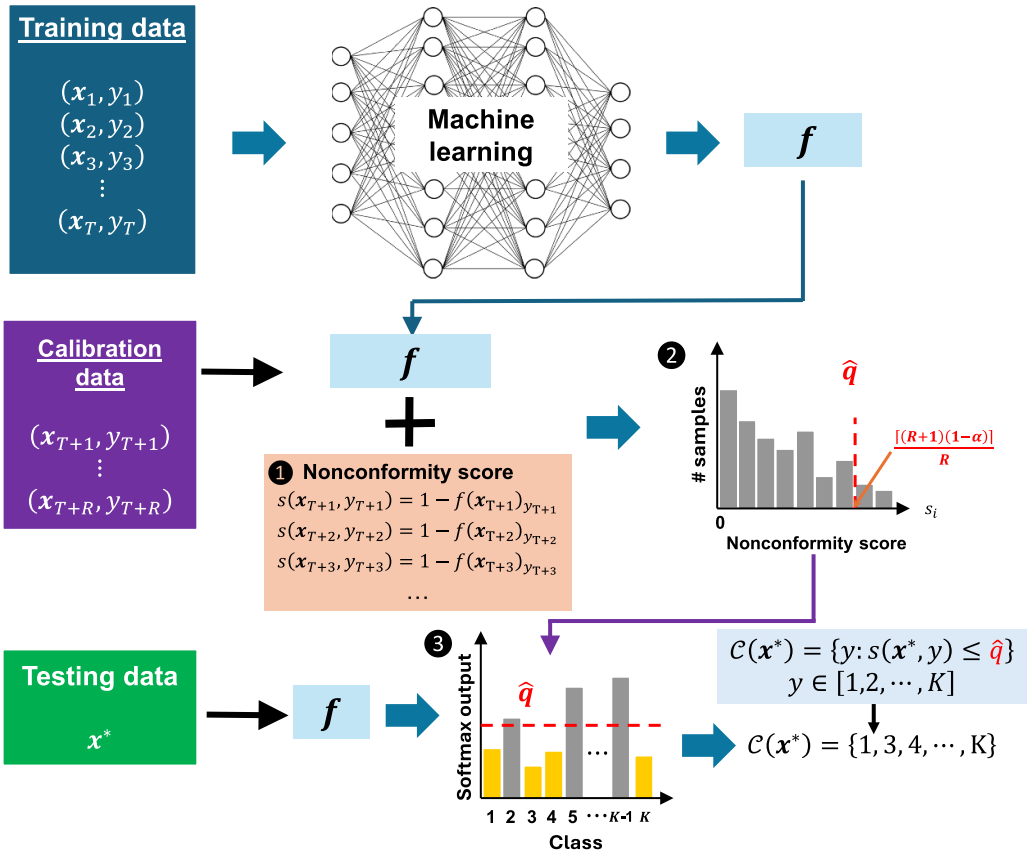


**Figure 4.** Procedures of conformal prediction in the context of $K$-class classification problem.

statistically valid uncertainty estimates without the need of calibration. However, the validity of conformal prediction is contingent upon data exchangeability while conformal prediction itself lacks the ability to detect data severely violating the condition of data exchangeability, such as OOD data. In this context, distance-aware uncertainty estimate, such as SNGP, demonstrates a desirable capability in distinguishing between in-distribution and OOD data.

### *3.2. Failure prediction*

UQ measures the confidence of model prediction by developing a probabilistic counterpart to the deterministic neural network. The built-in uncertainty estimate of the probabilistic model is then used to distinguish low-confidence vs high-confidence predictions. Unlike UQ, two separate models are developed in failure predictions: a base model and a meta model [90]. The base model is trained to perform certain classification or regression task, while the meta model is trained to estimate the confidence of the base model in failing/succeeding at the task. At a high level, the trained meta model serves as an 'observer' on top of an existing base model and it employs the estimated confidence to detect the prediction failures of the base model. Theoretically, the idea of failure prediction is analogous to the concept of stacked generalization [91]. Empirically, Blatz *et al* [92] showed that training a separate model achieved better performance in confidence estimation than solely relying on the original base model. In failure predictions, the base model and the meta model typically share the same feature representations derived from the encoder of the base model.

In the literature, a series of approaches have been developed to leverage the concept of failure prediction for creating reliable AI systems. For example, Corbière *et al* [66, 93] proposed using TCP (e.g. probability of the trained classifier in labeling $x$ as the correct class $y$) rather than the MCP as a target criterion for representing model confidence and developed an auxiliary model to learn TCP based on the feature representations extracted from the base model for failure predictions. Tsiligkaridis [94, 95] utilized uncertainty-aware deep Dirichlet neural networks to separate the confidence of correct and incorrect predictions and used TCP for identifying overconfident but incorrect model predictions. Adomavicius and Wang [30] proposed using the estimated absolute prediction error as the indicator of individual prediction reliability and recast the reliability estimation of individual predictions as a canonical numeric prediction problem. Following this idea, Zhang and Bose [29] defined the reliability of a ML model with respect to its individual prediction as the probability of the observed difference between the prediction of ML model and the actual observation falling within a small interval when the input varies within a small range subject to

**Table 1.** Quantitative comparisons of UQ methods for deep neural networks.

| Method | MCMC | Variational inference | Monte Carlo dropout | Deep ensemble | SNGP | Conformal prediction |
|---|---|---|---|---|---|---|
| Changes to the deterministic neural network for UQ | High | High | Low | Low | Medium | None |
| Scalability in model training for UQ | Low | Low | High | Medium | High | Not applicable |
| Is the uncertainty estimate statistically meaningful? | ✘ | ✘ | ✘ | ✘ | ✘ | ✓ |
| Need calibration after UQ? | ✓ | ✓ | ✓ | ✓ | ✓ | ✘ |
| Validity of UQ guaranteed after calibration? | ✘ | ✘ | ✘ | ✘ | ✘ | ✓ |
| Computational efficiency for UQ at inference stage | Low | Low | Medium | Medium | High | High |
| Distance-aware uncertainty estimate? | ✘ | ✘ | ✘ | ✘ | ✓ | ✘ |
| Model agnostic? | ✘ | ✘ | ✓ | ✓ | ✓ | ✓ |
| Performance in OOD detection | Low | Low | Low | Medium | High | None |
| Discrimination power of uncertainty-based confidence | Low | Low | Low | Medium | Medium | High |

a preset distance constraint. Based upon this definition, they developed a two-stage ML-based framework to directly learn the relationship between the input and the corresponding reliability estimate, thus providing an essential layer of safety net for adopting ML models in risk-sensitive environments.

In addition to the aforementioned methods, recent studies have started to investigate the learning algorithm and training data for developing more accurate model for failure prediction. In this regard, the research group from the National Laboratory of Pattern Recognition at the Chinese Academy of Sciences has made seminal contributions [100]. For example, Zhu *et al* [65, 96] revealed that prevailing confidence calibration methods often resulted in poor discrimination between the confidence of correct and incorrect predictions and were not practically useful for deciding whether to trust a prediction or not. To address this problem, they developed flat minima for increasing the discrimination power of the estimated confidence. Since the misclassification of neural network is a low-probability event, Zhu *et al* [97] mixed in-distribution data with outlier data to create synthetic samples and associated these synthetically generated samples with soft labels to increase the confidence separability between correctly classified and misclassified samples. Later, Zhu *et al* [101] developed a unified framework to perform OOD and misclassification detection simultaneously and leveraged sequence learning to fine-tune any given model through reliable weight consolidation and weight space interpolation for reliable misclassification and OOD detection. Grabinski *et al* [102] performed an extensive study on the confidence estimate of adversarially trained robust models in the literature and found out that non-robust models were overconfident with their false predictions. Interestingly, Rabanser *et al* [98] developed the first framework for failure predictions from the perspective of neural network training dynamics and used the disagreement of intermediate models with the final predicted label obtained during model training to signal the failure predictions of neural network. Liu *et al* [99] investigated the underlying mechanisms of hallucination in vision-language models (VLM) and attributed hallucination of VLM in part to the sensitivity of text encoders to vision inputs. To fix this, they developed visual and textual intervention to reduce VLM hallucinations by steering latent space representations during inference to enhance the stability of vision features. Recently, Zhang *et al* [103] conducted a comprehensive review of learning to reject, highlighting failure prediction as an important approach to implementing this idea.

**Table 2.** Comparisons of failure prediction methods for deep neural networks.

| Literature | Method | Focused aspect | Mechanism |
|---|---|---|---|
| Corbière *et al* [66, 93] | Meta model for TCP prediction | Confidence measure for failure prediction | Confidence-based failure/success discrimination |
| Tsiligkaridis [94, 95] | Information-robust Dirichlet (IAD) network for confidence estimation | IAD for better TCP prediction | Separate correct and incorrect model predictions using the TCP metric |
| Adomavicius and Wang [30] | Meta model for predicting the absolute prediction error of the base model | Model reliability for individual prediction | Meta model to estimate the prediction error of the base model |
| Zhang and Bose [29] | Development of model reliability measure for individual prediction | Meta model for predicting model reliability specific to individual prediction | Develop better reliability indicator for informed failure prediction |
| Zhu *et al* [65, 96] | Flat minima for reliable confidence estimation | Model training and learning algorithm | Enlarge the confidence gap between correctly and incorrectly classified samples |
| Zhu *et al* [97] | Generate synthetic data via linear interpolation to reinforce the signal of misclassified samples | Training data augmentation | Augment the signal of misclassified samples to increase the exposure of low-density region |
| Rabanser *et al* [98] | Use the agreement of the predictions between intermediate models and final model prediction as a means for failure prediction | Training dynamics | Correct neural network predictions exhibit better agreement with the final prediction label and vice versa |
| Liu *et al* [99] | Steer the latent space representations to enhance the stability of vision features during model inference | Causes to hallucinations of VLM | Reduce the sensitivity of textual encoder to vision inputs by enhancing the stability of vision features |

Table 2 summarizes the engineering mechanisms of existing methods for failure prediction of deep neural networks. As can be seen, most methods focus on establishing informative confidence indicators for failure predictions of neural networks. A limited number of methods have attempted to identify cues and precursors signaling failures of neural network at its predictions. For example, Rabanser *et al* [98] examined the training dynamics of neural network and discovered that the neural network prediction failures tended to exhibit a high prediction instability across intermediate model states obtained during model training. A few methods concentrate on improving the learning and training algorithms to enhance the discrimination capability of confidence estimate. From the information in table 2, we notice that predicting neural network failures is still in its early stages of development. Yet the literature still lacks a comprehensive and holistic method for reliable failure predictions of neural networks.

### 3.3. Learning with abstention

Another prevalent strategy to build models that recognize their own limitations is through learning with abstention (or known as selective prediction). In fact, the concept of learning with a reject option was already investigated by Chow over 60 years ago [104, 105]. In the literature, there are two popular means to implement learning with rejection: designing a cost function that assigns varying costs to different types of errors (e.g. wrong predictions, abstention) to encourage neural network to learn when to abstain; developing a learnable function to serve as proxies of confidence estimation for model abstention. Regardless of differences in the technical details, these methods are devised to allow the model to refrain from making predictions whenever they are not sufficiently confident in their predictions.

*3.3.1. Cost-sensitive machine learning.* Cost function-based approaches for learning with abstention, also known as cost-sensitive machine learning, assign varying costs to different types of errors [106–108]. These assigned costs represent the penalties attributed to each rejected and misclassified sample, as well as the gain associated with each correctly classified sample. Note that incorrect predictions are often assigned a higher cost than abstention. By designing the cost matrix in this way, we aim to encourage the model to abstain when a particular input is difficult to classify, rather than forcing it to make inaccurate prediction. After the cost matrix is designed, deep learning models are trained to minimize the expected overall cost by optimizing when to refrain from making predictions, when to make predictions. Along this direction, De Stefano *et al* [109] were among the first few to investigate learning with rejection in the context of neural networks, and developed an effectiveness function to measure the utility of the reject option tailored to a considered application domain. Charoenphakde *et al* [110] proposed a

surrogate loss-based approach to learn an ensemble of cost-sensitive classifiers to approach multi-class classification with rejection. Nguyen and Hullermeier [111] developed a formal framework for multi-label classification with partial abstention by extending an underlying multi-label classification loss function to accommodate abstention in the learner. Kalai and Kanade [112] developed an approach for optimally abstaining in classification and regression in the transductive setting to avoid making predictions on 'blind spot' due to distributional shift or adversarial examples.

*3.3.2. Selective prediction.* Another way to implement learning with a reject option is to integrate the learning of an abstention function into the standard training workflow of neural network. The trained abstention function is then used to inform when to accept or defer neural network predictions. Typically, we add a binary classifier in the output layer of neural network to indicate whether to abstain or not. Along this direction, Geifman and El-Yaniv [113] proposed to construct a selective classifier $(f, g)$ that would guarantee a desired error rate with a high probability, where $f$ was the standard neural network classifier and $g$ was a rejection function built upon the softmax response or prediction uncertainty estimated by MC dropout. To combat label noise, Thulasidasan *et al* [114, 115] extended the standard cross-entropy training loss to accommodate the abstention option. The incorporation of abstention allows neural network to abstain on confusing samples while continuing to learn and improve its classification performance on the non-abstained samples. Subsequently, Geifman and El-Yaniv [116] developed SelectiveNet, a multi-headed neural network for end-to-end learning of classification (or regression) and rejection simultaneously to satisfy a required coverage. Recently, Guo *et al* [117] curated a challenging UNK-VQA dataset to probe the capability of several multi-modal visual question answering (VQA) large models in refraining from answering questions that cannot be answered or were beyond their scope of knowledge. To address this problem, they developed a selective classifier with an integrated confidence function to control the overall model prediction versus abstention level.

Clearly, cost function-based machine learning and learning with an integrated reject option represent two different engineering mechanisms for implementing abstention in the neural network. The former uses a cost function to steer neural network in learning when to abstain or make predictions. In contrast, the latter incorporates the abstention option directly into the learning process and aims to strike a balance between model coverage and model risk. Table 3 summarizes existing engineering strategies for learning with abstentions. It can be observed that most cost-sensitive learning methods focus on establishing optimal rejection rules for a variety of use scenarios, either through analytical approaches or by using customized loss functions. Unlike cost-sensitive learning methods, selective prediction seeks to integrate the learning of abstention and classification using a specialized neural network architecture for bounding model risk. Despite their difference, these two engineering paradigms target for reducing model risk by abstaining from difficult samples for approaching functionally reliable AI systems.

### 3.4. Formal verification of neural networks

Formal methods refers to a class of logic and mathematics techniques, including model checking, deductive verification, integer programming, to provide a rigorous guarantee about the correctness of computer software [118–121]. Formal verification of a neural network is the process of mathematically proving (or finding a counterexample to) a specified property (such as safety or robustness) for all possible inputs within a given domain. Formal verification is often a must step towards satisfying the stringent safety assurance requirements mandated by international standards for software embedded in mission-critical systems, such as AI-based diagnosis, autonomous driving, aircraft auto-piloting. Since verifying the properties of neural networks formally is a challenging task, existing studies attempt to explore the verification of neural network via simplification and approximations. By formally verifying the behavior of neural network, we hope to prove that neural networks have some desired properties and behavior in robustness, safety, and correctness we can formally trust.

In the case of a simple binary classification, consider an input $\boldsymbol{x}_0$, assume $f(\boldsymbol{x}_0) > 0$, robustness verification of neural network is defined as verifying the property of $f(\boldsymbol{x}) > 0$, $\forall \boldsymbol{x} \in \mathcal{C}$, where $\mathcal{C} := \left\{ \boldsymbol{x} | \left\| \boldsymbol{x} - \boldsymbol{x}_0 \right\|_p \leqslant \epsilon \right\}$ is a $l_p$ norm ball defining a perturbation set around $\boldsymbol{x}$. Ideally, we hope that a small perturbation to the input $\boldsymbol{x}_0$ should not result in changes to the output of the neural network. This is why it is important to equip neural networks with the robustness property. In essence, the idea of formal verification is to encode the neural network and the property we are interested in verifying as a formal statement, using integer linear programming (ILP), satisfiability modulo theory (SMT) or Boolean satisfiability problem (SAT). Take SMT as an example. It is a set of concrete tools, automated theorem provers that extend the SAT to more complex formulas involving real numbers, integers, and/or various data structures, and apply specialized decision procedures to answer the question 'is this formula ever true?'. In practice, an SMT solver can encode a neural network's arithmetic as logical constraints and attempt to find an input that violates the property (a counterexample). Besides robustness, formal verification can also target other properties of neural network, such as safety, correctness. For instance, we can formally verify neural network's compliance with application-specific safety constraints (e.g. a robot arm never enters an unsafe region), monotonicity or fairness properties, or bounded error in regression outputs, depending on the specification.

As formally verifying the properties of ReLU neural networks involving non-convex constraints is NP-complete, extant studies have explored the verification of neural

**Table 3.** Comparisons of learning with abstention methods.

| Category | Literature | Method | Mechanism |
|---|---|---|---|
| Cost-sensitive learning methods | Chow [104, 105] | Analytical solution to establish optimum rejection rule | Strike a tradeoff relation between model error and rejection for Bayes optimum classifier |
| | De Stefano *et al* [109] | Defined an effectiveness function to measure the utility of abstention in neural networks | Uncover root cause of low-reliability predictions and establish customized optimal thresholds for abstention in each case |
| | Charoenphakde *et al* [110] | Develop a surrogate loss for cost-sensitive learning approach for classification with rejection | Establish an optimal rejection rule for multi-class classification |
| | Nguyen and Hullermeier [111] | Partial abstention for multi-label classification | Extend the classification loss function to accommodate the abstention option for partial rejection |
| | Kalai and Kanade [112] | Transductive algorithm for abstaining from prediction with OOD test examples | Information-theoretic formulation for model abstention in the presence of adversarial test examples and covariate shift |
| Selective prediction | Geifman and El-Yaniv [113] | Selective classifier for guaranteed risk control | Learning of selective classifier for guaranteed risk control |
| | Thulasidasan *et al* [114, 115] | Deep abstaining classifier for robust learning in the presence of label noise | Extend cross-entropy loss for incorporating the learning of abstaining on noisy data |
| | Geifman and El-Yaniv [116] | SelectiveNet for integrated learning of abstention and classification | Customized neural network architecture and loss function for integrated learning of classification and abstention |
| | Guo *et al* [117] | Dedicated dataset for probing the abstention ability of foundational VQA model | Apply selective prediction in the context of large VQA model |

networks via simplifications and linear relaxations to approximate neural network's decision boundary [126–128]. For example, Pulina *et al* [122] combined counterexample-triggered abstraction refinement with SMT to verify the safety of multi-layer perceptron, but the developed method was limited to small-size neural networks. To simplify formal verification, many studies consider networks with nodes having piecewise linear activation functions as they are more amenable to formal verification. For example, Huang *et al* [123] proposed an automated verification framework for feed-forward multi-layer neural networks based on SMT, where the nodes of neural networks were assumed to have piecewise linear activation functions. Ehlers [118] considered the type of feed-forward neural network with piece-wise linear activation functions and generated a linear approximation of the overall network behavior that can be added to SMT or ILP instances for encoding neural network verification problems. Wang *et al* [129] developed Neurify to efficiently verify the safety properties of neural networks and identified concrete counterexamples to demonstrate the violations of safety properties. Sun *et al* [124] utilized a satisfiability modulo convex encoding to list the possible assignments of different ReLUs to verify the safety of a NN controller in yielding control actions for an autonomous robot. Venzke and Chatzivasileiadis [125] proposed the first framework based upon mixed-ILP to build formal guarantees of neural network behavior for power system applications, and they formally proved that no adversarial examples can exist for a continuous range of neural network inputs. For a thorough survey on formal methods applied to deep learning, refer to Urban and Miné [130].

Table 4 summarizes formal verification methods for neural networks in the literature. Due to the lack of a widely accepted, accurate, and scalable mathematical specification for characterizing the behavior of neural network, existing studies often focus on verifying neural networks against simple input–output specifications by making certain assumptions on the activation functions. Importantly, these formal verification methods are only applicable to small-scale neural network with simplified network architecture and task-specific functions. Yet no method has been developed to formally verify the behavior of large-scale neural network (e.g. foundation models with versatile functions) up to now. Nevertheless, formal verification contributes to AI reliability by providing strong guarantees of performance within specified conditions, thereby addressing certain failure modes that testing or probabilistic methods might miss.

**Table 4.** Comparisons of formal verification methods for neural network.

| Literature | Method | Assumption | Mechanism |
|---|---|---|---|
| Pulina *et al* [122] | Abstractions of multi-layer perceptrons to corresponding Boolean combinations of linear arithmetic constraints | The activation function of the hidden layer neurons follows the logistic function | Generate spurious counterexamples to automate the correction of misbehaviors of abstracted MLP representations |
| Huang *et al* [123] | SMT-based automated verification framework for feed-forward multi-layer neural networks in image classification | Local semantic constancy, invariance under defined manipulations, sufficiency of discretisation, and soundness of propagation. | SMT-based search of adversarial examples by discretization |
| Ehlers [118] | A global linear approximation of the overall network behavior that can be added to SMT or integer linear programming instances encoding neural network verification problems | Nodes follow piece-wise linear activation, linear/convex specifications over network inputs and outputs and known input bounds. | Linear approximation of the overall neural network behavior |
| Sun *et al* [124] | Formal verification of the safety of an autonomous robot equipped with a neural network controller for processing LiDAR images to produce control actions | Linear robot dynamics, environment is polytopic, ReLU-only neural networks as controller, and a fixed LiDAR heading. | A finite state abstraction of the system and standard reachability analysis over the finite state abstraction to compute the set of safe initial states |
| Venzke and Chatzivasileiadis [125] | Mixed-integer linear programming for formal guarantee of neural network behavior in power system applications | ReLU is selected as the activation function in the neural network | Reformulation of ReLU as a mixed-integer program |

### *3.5. Knowledge-enabled AI*

Although deep learning exhibits a promising performance, the purely data-driven model might suffer from poor generalization and generates predictions violating well-established knowledge (e.g. physical laws, guidelines). To address these problems, researchers have explored to integrate prior knowledge in various forms, such as differential equations, logic rules, algebraic equations, causal graphs, knowledge graphs, into the learning process of neural networks to guide its training [131]. Integration of domain knowledge serves to induce neural networks to learn the right representations for reasoning and decision making. In essence, the injected domain knowledge serves as inductive biases on top of the observational ones to regularize the learning of neural network. The combination of model parameters leading to predictions violating the well-established domain knowledge are often penalized during the model training. In this section, we take physics-informed neural network (PINN) and causality-informed neural network as two examples to illustrate the knowledge-enabled AI paradigm.

PINN is a representative example of the knowledge-enabled AI computational paradigm as it encodes mathematical physics models into neural network [132, 133]. Given the broad scope of physics knowledge, there are various means to incorporate physics knowledge into neural network, such as physics-based constraints, loss function, PINN architecture. In most PINN studies [134–136], physics-based loss terms are formulated to characterize the losses associated with the initial and boundary conditions of PDE as well as the residual of approximating PDE equations. In addition to the loss function-based approach, there are also other means to inject physics into the learning systems [137], such as data augmentation, transfer learning, delta learning. For example, in data augmentation, synthetic data extracted from first-principle simulations representing physical knowledge is combined with actual observation data to build an augmented dataset to train deep learning models. Besides, physics knowledge has also been utilized to inform the architecture design of neural networks. This line of research seeks to incorporate the physics properties into nodes or layers of neural networks to make the black-box algorithm more interpretable and generalizable. In other words, by associating neurons/layers with physical equations, we aim to ensure that these nodes and layers generate physically consistent results in the neural network. For example, Yucesan *et al* [138] combined estimation of bearing fatigue damage increment through physics-informed kernels with estimation of grease damage increment through a multi-layer perceptron to characterize bearing fatigue and grease damage accumulation.

Similarly, there have been extensive studies exploring the utilization of causal knowledge to inform the learning of neural network [139]. For example, Kyono *et al* [140] proposed to leverage causal directed graph learning as an auxiliary task to regularize the training of neural network. In fact, the joint learning of causal directed graph learning embedded in neural network training acts as a feature selection regularizer to shrink the weights of non-causal predictors, thereby facilitating the discovery and reliance towards causal predictors. Teshima *et al* [141] developed a model-agnostic data

**Table 5.** Comparisons of knowledge-enabled AI methods.

| Literature | Method | Assumption | Mechanism |
|---|---|---|---|
| Karniadakis *et al* [132] | Physics-informed neural network | Physical equations characterizing the behavior of the system of interest are available | Encode governing laws in the form of physical equations (e.g. PDE, ODE) into the loss function of neural network |
| de Beaulieu *et al* [148] | Physics-informed degraded data augmentation | Physics of failures about the system are known | Generate additional data mimicking physics of failures to enhance model training |
| Kapusuzoglu and Mahadevan [149] | Pretraining of ML model using physics-informed synthetic data | Physical simulation is available | Use physics-informed synthetic data to first train the ML model and update it with experimental data |
| Kyono *et al* [140] | Causal DAG-regularized loss function | Causal DAG follows linear relationships | Induce neural network to focus on learning with stable and meaningful causal predictors |
| Teshima *et al* [141] | Augment the training data based on conditional independence in causal graph for supervised machine learning | Conditional independence in causal graph is available | Augment training data with samples representing conditional independence in the causal graph |
| Kancheti *et al* [142] | Domain prior on causal relationships | Prior knowledge on causal relationship is known and stable | Incorporate domain prior as a regularizer in the learning of neural network |
| Zhai *et al* [146] | Learn causal DAG from observational data | Causal relationships are linear in the DAG | Learning with causal feature representation in graph neural network for click-through rate prediction |

augmentation method to leverage causal prior knowledge of conditional independence relations encoded in a causal graph to improve the generalization of machine learning models. Kancheti *et al* [142] proposed a regularization method to align the learned causal effects of a neural network with domain priors, including both direct and total causal effects, for enhancing the model robustness and accuracy when dealing with noisy data. Wen *et al* [143] considered inter-variable causal relations and developed a causally-aware generator with a tailored architecture to generate synthetic data that can capture target data distribution more accurately. Cui *et al* [144, 145] formulated a stable learning paradigm in an attempt to establish a common ground between causal inference and machine learning for enhancing model generalization in unseen environments. Zhai *et al* [146] developed a causality-based click-through rate prediction model in the graph neural networks (GNNs) framework, where causal relationships among field features were discovered and retained in GNNs via a structured representation learning approach. Recently, Berrevoets *et al* [147] outlined the roadmap for the development of causal deep learning framework spanning structural, parametric, and temporal dimensions to unlock its potential in solving real-world problems.

The preceding two paragraphs illustrate how physical and causal knowledge can be integrated into neural networks and their role in enhancing the generalization and reliability of AI algorithm. Table 5 summarizes several knowledge-enabled AI methods in the literature. It can be observed that knowledge is manifested in various forms, such as differential equations, conditional independence, causal relationships, causal effect. Existing studies have used domain knowledge to regularize the learning process of neural network by either changing the loss function or generating synthetic data representing the prior knowledge. Incorporating domain knowledge undoubtedly improves the stability of neural network in performance. However, current engineering practices cannot ensure that the neural network consistently adheres to the established domain knowledge. Additional research is needed to provide provable guarantees of compliance with domain knowledge.

Note that our coverage on knowledge-enabled AI is not intended to be exhaustive; rather, it aims to establish the connection between knowledge-enabled AI and AI reliability. Importantly, in addition to integrating domain knowledge into AI models to enhance their generalization, domain knowledge can also be utilized to empirically evaluate the reliability of AI systems. Established domain knowledge could be leveraged to inform the design of a suite of targeted test cases (e.g. edge cases, corner cases, boundary conditions, design of experiments) for assessing the reliability of AI systems in the intended operating environment and evaluating their compliance with safety and reliability standards and certification requirements. For example, Hager *et al* [19] created a curated dataset based on the Medical Information Mart for Intensive Care (MIMIC-IV) database spanning 2400 real patient cases and four common abdominal pathologies to evaluate the compliance of LLMs with diagnostic and treatment guidelines in medical practice. These test cases informed by domain knowledge will provide evidence of the reliability of the AI model in different contexts and offer insights on the direction

to improve AI reliability. Pei *et al* [150] incorporated custom domain-specific constraints to generate valid and realistic instances for systematically testing and exposing erroneous corner case behaviors of real-world deep learning systems.

## 4. Research challenges and opportunities

As previously discussed, the fields of reliability engineering, risk management, and trustworthiness assurance for AI systems are still in their infancy. However, there is a rapidly growing demand for understanding the behavior of AI models, such as how AI model performance degrades, how AI model exhibits its behavior that is physically inconsistent or implausible, how AI model leads to mishap events. Understanding the failure modes of AI models is crucial to the subsequent development of fault diagnosis and model revision methods for preventing the recurrence of the same error in the future. In this section, we highlight several research challenges in the study of reliability engineering and trustworthiness assurance for AI systems.

(i) **Unclear failure modes**. To the best of our knowledge, we still do not fully understand when a well-trained AI model will fail and how it produces erroneous predictions. The lack of clear understanding on the failure mechanisms of AI have become as a major roadblock to the progressive development of reliable AI systems in the long run. Traditionally, when encountering a bug in the software systems, developers address the issue by directly troubleshooting and correcting the source codes. However, in the context of AI systems, traditional software debugging methods are no longer applicable because the logic within AI models is not explicitly programmed but learned from data. Furthermore, with millions or even billions of parameters, it is challenging to pinpoint which neurons are responsible for the erroneous behavior, let alone determine which parameters need adjustment when the model makes a mistake. Although data augmentation or model retraining can be used to fix some errors, retraining is unfortunately expensive and time-consuming while they provide no guarantee on fixing the misbehavior of AI model.

One possible remedy is to employ explainable AI (XAI) methods in the literature [31, 151, 152], such as LIME [39], Shapley Value [153], saliency map [154, 155], to uncover the reasoning process of AI models for the purpose of diagnosing and troubleshooting AI system misbehavior. Unfortunately, XAI is often performed on an instance-by-instance basis [156, 157]. While this approach offers valuable insights tailored to individual cases or scenarios, instance-level XAI does not provide a comprehensive understanding of model-level behavior while such analytics are essential for effective troubleshooting. More importantly, XAI are not standardized or consistent across cases, this in turn limits the development of advanced tools for automating model diagnosis and troubleshooting with XAI. Furthermore, many existing XAI techniques are designed primarily to elucidate the decision-making process of AI models, often focusing on explanation for its own sake. However, the explanations they provide may not accurately reflect the true reasoning mechanisms of the original model. Even if we set these issues aside, XAI still needs significant development to identify the specific neurons responsible for the surfaced misbehavior of the AI systems. Right now, we still lack a comprehensive understanding of how neurons orchestrate within neural networks, making it even more challenging to locate the root causes of model misbehavior.

(ii) **No theoretical guarantee on the performance of detecting input data breaking i.i.d condition in the open world**. As environment plays a crucial role in the reliability engineering of AI models, there is currently no method with a provable performance guarantee on the detection of input data violating i.i.d. that is foundational to deep learning. The lack of guarantee on the validity of the deployment environment poses a significant risk for AI systems to function reliably in practice. While formal methods offer provable guarantee on the robustness of neural network, they are restricted to small-size neural network by introducing relaxations on the activation functions of neural network. Since the deep learning model is built upon the i.i.d. foundation, the resulting model is limited in its capability to handle inputs from a distribution different than the training data. The unknown situations (e.g. OOD; covariate shift) arising from the model input pose huge risks and might lead the model to behave unexpectedly. As both dataset shift and OOD data fundamentally breaks the i.i.d condition crucial for the reliable operation of AI systems, these AI models often experience severe performance degradation and generate misleading predictions when dealing with these data different from the training data. Although some progress has been made to detect input data breaking the i.i.d. condition, there is no theoretical guarantee on the performance of these methods in detecting the unsafe input data. Therefore, how to create a valid environment for AI models to operate safely and reliably remains an open issue to be addressed.

(iii) **Lack of an ecosystem for the progressive development of reliable AI.** When an AI model exhibits misleading behavior, we often take a piecemeal approach to fix the misleading behavior by either model retraining or data augmentation. Although there are already quite a number of studies on interpretable or XAI, most of these studies concentrate on visualizing the saliency map for associating the model prediction with the corresponding evidence in the input data while they provide limited value in uncovering the layer-by-layer reasoning process of neural network [158–160]. To the best of our knowledge, a holistic approach is still lacking in the literature for the progressive development of reliable AI. How to develop a comprehensive ecosystem to provide one-stop service for reliability-related analysis, including

model behavior analysis, misbehavior attribution, model reliability analysis, blind spot identification, fault diagnosis, model update, model editing, training data probing, is highly important to the sustainable development of reliable and trustworthy AI in the long term. In addition, it is also important to create benchmark datasets to facilitate reliability studies and model development along these fronts. Establishment of relevant methods and data for benchmark purpose will foster the creation of a growing research environment and community for reliability studies of AI.

These research challenges also present valuable opportunities to researchers in the filed of computer science, data science, statistics, as well as reliability and risk engineering. Since there is an urgent need for developing reliability modeling and trustworthiness assurance methods that address the unique characteristics of AI-related risks, there are plenty of opportunities along the development of reliable and trustworthy AI. Researchers could contribute to the development of reliable and trustworthy AI from the following aspects:

(i) **Development of an effective and scalable framework for troubleshooting and model revision to progressively enhance the performance of AI systems**. A major obstacle to the progressive development of reliable and trustworthy AI is the lack of a mature framework for diagnosing the system's performance and correcting the model/data error when the system behaves erroneously. In reliability engineering, numerous methods have been developed to diagnose the faults over a wide range of advanced equipment and complex engineering systems, such as aircraft, nuclear power plant, helicopter, power grid. It is worthwhile to explore how we can exploit existing techniques in fault diagnosis, such as fault detection, fault isolation, fault identification, root cause analysis, fault evaluation, corrective actions, and apply troubleshooting techniques accumulated from traditional engineering systems to understanding and analyzing the behavior of AI-powered intelligent systems. In this regard, existing fault diagnosis methods need to be strengthened in efficiency and scalability to deal with the high-dimensional state space and massive number of parameters of AI systems. For example, how to implement and scale up Bayesian methods for troubleshooting the behavior of deep neural network. In addition, once the specific model parameters that need adjustment when the model makes a mistake are accurately identified, the next important step is to develop methods to update AI models in a way that is not as computationally expensive as training a new model from scratch. This is particularly important for LLMs as they require regular updates to stay current with the latest world knowledge. If we could develop a mature fault diagnosis and model revision method, we can improve the performance of AI model by fixing its errors and reducing its 'blind spot' step by step. Importantly, it is crucial to study measures to ensure that the fixed model will not commit the same error again in the future. By establishing such a progressive development scheme, we could gradually improve the reliability of AI models by lowering its error rate over time.

(ii) **Development of holistic methods for creating reliable and trustworthy AI systems**. Existing studies focus on a particular aspect of AI reliability, such as OOD detection, dataset shift; model uncertainty; integration of domain knowledge into neural network; data-centric AI. Clearly, a holistic approach is still missing for combining the collective strength of existing methods to form a comprehensive solution for safeguarding AI as a whole after its deployment in the open world. Given that no existing method can provide a theoretical guarantee of data or model trustworthiness in an open-world setting, it is valuable to apply reliability principles by combining several weak models to achieve a desired level of reliability, with each model complementing the strengths of the others. Importantly, this concept can also be expanded to leverage multiple AI agents or AI models with varying capabilities working collaboratively to achieve reliable and trustworthy AI. The collaborative approach is anticipated to harness complementarity-driven deferral by routing cases to the most suitable AI agent or model for dependable inference and reasoning, thereby leveraging their combined strengths and mitigating the blind spots of individual models.

Importantly, since data trustworthiness and model trustworthiness are coupled in a way that is not fully understood yet, it is valuable to perform in-depth investigations for uncovering their coupling relationships, such as analyzing the impact of untrustworthy data on the model trustworthiness, how to distinguish the effects on AI systems resulting from data trustworthiness versus model trustworthiness, methods for attributing erroneous predictions from AI systems to data and model. As discussed by Liang *et al* [40], creating appropriate data pipelines has become a major bottleneck in developing trustworthy AI algorithms due to the increasing maturity of model-building techniques. Noways, the role of data in training or evaluating AI is often sparsely discussed in the literature. Yet we lack a systematic approach to assess the impact of data cleaning, synthesizing, annotating, and valuing on creating reliable and trustworthy AI. Besides, in the case that if we have a module to detect AI's potential failures, it is helpful to embed safety-preserving fallback mechanism that is independent of AI model for staying on the safe side in critical applications [161].

(iii) **Development of resilient AI systems via efficient software testing and continuous human-AI collaborations.** Although AI systems can be very accurate for a specific task, it cannot be perfect (i.e. 100% correct). In other words, mistakes from AI systems are inevitable. In this circumstance, the key challenge lies in how to respond if the underlying AI system malfunctions. Addressing this challenge boils down to answering two important questions: under what scenarios that (or when to anticipate) the AI system is likely to be incorrect and what actions to

take if the AI system malfunctions. Application-specific domain knowledge and established testing methods in software development could be leveraged to design a comprehensive suite of test cases that efficiently and scalably expose the vulnerability of the underlying AI systems such as undetected OOD data and incorrect predictions on in-distribution input data. These identified failure cases or detected blind spots can then be stored in a database for future query and reference. During model deployment, if the AI system encounters inputs with characteristics similar to those in the database, these cases can either be deferred to human for further judgment or routed to backup systems capable of reliably handling these cases to ensure safety. This concept is similar to retrieval augmented generation (RAG) used in LLMs. Essentially, RAG reduces hallucinations of LLMs and bolsters the AI system resilience through external augmentation [162]. In other words, instead of solely relying on AI systems to process all types of inputs, we can enhance their reliability by selectively handling certain cases differently. By establishing customized safety guardrails to manage the blind spots of AI systems, we will significantly enhance the resilience of the AI-powered intelligent systems by anticipating potential failure scenarios before they occur. In addition, the enumerated failure cases can be further analyzed to understand why the trained AI systems fail when handling them, thus providing directions for the further improvement of AI system performance.

## 5. Conclusion

In this paper, we contextualize reliability engineering for AI systems and highlight the unique characteristics in the risks associated with AI. Since the risks pertaining to AI have their unique characteristics, traditional reliability modeling and risk management methods fundamentally lose their efficiency in characterizing and modeling AI system's risk. Hence, there is a pressing need for developing reliability engineering and risk management approaches that are dedicated to AI systems. In this paper, we systematically review several prevalent strategies in the extant literature to approach reliable and trustworthy AI, including UQ, failure prediction, learning with abstention, formal verification, and knowledge-enabled AI. Despite these rapid progress, the literature still lacks a comprehensive approach to support systematic reliability-oriented researches and studies surrounding AI systems ranging from misbehavior diagnosis to model editing. To tackle this problem, we elaborate on several research challenges in the reliability study of AI systems, and outline possible research opportunities that reliability researchers could contribute to this research topic with pivotal importance. We hope that this work will motivate researchers to develop rigorous methods to manage the diverse sources of risks for creating reliable and trustworthy AI systems such that we could use AI to empower critical applications in an responsible and orderly manner.

Note that this review paper primarily focuses on AI systems in the realm of supervised learning and has not considered more challenging reliability problems that might arise when AI systems interact with other modules within an application. For example, how the AI-powered autonomous driving system is compromised when onboard sensor feeding AI systems malfunctions, how the prediction error of AI systems propagates, affects the precision of control actions, and leads to an undesirable self-driving experience. In addition, even though we could establish an objective reliable AI systems, AI trustworthiness also involves user perceptions, transparency, and interpretability. This review paper has not considered how end-users, particularly in high-stakes settings, interact with AI systems and perceive the established AI system reliability. Human-in-the-loop experimental study is needed for ensuring the alignment between human confidence and AI reliability.

## Acknowledgment

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25206422), the Research Committee of The Hong Kong Polytechnic University (Project code: RNAH), the National Natural Science Foundation of China (Grant No. 62406269), the JST CRONOS Grant (No. JPMJCS24K8), and the JSPS KAKENHI Grant (No. JP21H04877, No. JP23H03372, and No.JP24K02920).

## ORCID iD

Tao Wang https://orcid.org/0000-0003-0201-8100

## References

[1] Goodfellow I 2016 *Deep Learning* (MIT Press)
[2] Zio E 2022 Prognostics and health management (PHM): where are we and where do we (need to) go in theory and practice *Reliab. Eng. Syst. Saf.* **218** 108119
[3] Nemani V, Biggio L, Huan X, Hu Z, Fink O, Tran A, Wang Y, Zhang X and Hu C 2023 Uncertainty quantification in machine learning for engineering design and health prognostics: a tutorial *Mech. Syst. Signal Process.* **205** 110796
[4] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[5] Heim E, Wright O and Shriver D 2025 A guide to failure in machine learning: reliability and robustness from foundations to practice (arXiv:2503.00563)
[6] Konstantopoulos S 2024 On the reliability of artificial intelligence systems *Proc. 13th Hellenic Conf. on Artificial Intelligence* pp 1–4
[7] Abdar M, Khosravi A, Islam S M S, Acharya U R and Vasilakos A V 2022 The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process *IEEE Syst. Man Cybern. Mag.* **8** 28–40
[8] Begoli E, Bhattacharya T and Kusnezov D 2019 The need for uncertainty quantification in machine-assisted medical decision making *Nat. Mach. Intell.* **1** 20–23

[9] Wang L, Lin L and Dinh N 2024 Trustworthiness modeling and evaluation for a nearly autonomous management and control system *Reliab. Eng. Syst. Saf.* **245** 110008

[10] Nsoesie E O and Ghassemi M 2024 Using labels to limit AI misuse in health *Nat. Comput. Sci.* **4** 638–40

[11] Kaur D, Uslu S, Rittichier K J and Durresi A 2022 Trustworthy artificial intelligence: a review *ACM Comput. Surv.* **55** 1–38

[12] Zheng S *et al* 2025 Bridging the data gap in AI reliability research and establishing DR-AIR, a comprehensive data repository for AI reliability (arXiv:2502.12386)

[13] Lambert B, Forbes F, Doyle S, Dehaene H and Dojat M 2024 Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis *Artif. Intell. Med.* **150** 102830

[14] Benjamens S, Dhunnoo P and Meskó B 2020 The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database *npj Digit. Med.* **3** 118

[15] Zhang X, Chan F T, Yan C and Bose I 2022 Towards risk-aware artificial intelligence and machine learning systems: an overview *Decis. Support Syst.* **159** 113800

[16] Zhou L, Schellaert W, Martínez-Plumed F, Moros-Daval Y, Ferri C and Hernández-Orallo J 2024 Larger and more instructable language models become less reliable *Nature* **634** 1–8

[17] Nguyen A, Yosinski J and Clune J 2015 Deep neural networks are easily fooled: high confidence predictions for unrecognizable images *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 427–36

[18] Majumder S, Dong L, Doudi F, Cai Y, Tian C, Kalathil D, Ding K, Thatte A A, Li N and Xie L 2024 Exploring the capabilities and limitations of large language models in the electric energy sector *Joule* **8** 1544–9

[19] Hager P *et al* 2024 Evaluation and mitigation of the limitations of large language models in clinical decision-making *Nat. Med.* **30** 2613–22

[20] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y J, Madotto A and Fung P 2023 Survey of hallucination in natural language generation *ACM Comput. Surv.* **55** 1–38

[21] Wang B *et al* 2023 DECODINGTRUST: a comprehensive assessment of trustworthiness in gpt models *Advances in Neural Information Processing Systems* vol 36

[22] Rawte V, Sheth A and Das A 2023 A survey of hallucination in large foundation models (arXiv:2309.05922)

[23] Liang S, Li Y and Srikant R 2018 Enhancing the reliability of out-of-distribution image detection in neural networks *6th Int. Conf. on Learning Representations, ICLR 2018*

[24] A driverless car hits a person crossing against the light in China 2024 (available at: https://apnews.com/article/chinaautonomous-driving-accident-baidu-b0b4527ff355836f2df03868ff0bd0fc) (Accessed 30 September 2024)

[25] Reliability Engineering 2024 (available at: https://en.wikipedia.org/wiki/Reliability_engineering#::text=Reliability%20engineering%20deals%20with%20the,achieved%20by%20mathematics%20and%20statistics) (Accessed 30 September 2024)

[26] Hong Y, Lian J, Xu L, Min J, Wang Y, Freeman L J and Deng X 2023 Statistical perspectives on reliability of artificial intelligence systems *Qual. Eng.* **35** 56–78

[27] Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M and Wichmann F A 2020 Shortcut learning in deep neural networks *Nat. Mach. Intell.* **2** 665–73

[28] Sinha R, Sharma A, Banerjee S, Lew T, Luo R, Richards S M, Sun Y, Schmerling E and Pavone M 2022 A system-level view on out-of-distribution data in robotics (arXiv:2212.14020)

[29] Zhang X and Bose I 2024 Reliability estimation for individual predictions in machine learning systems: a model reliability-based approach *Decis. Support Syst.* **186** 114305

[30] Adomavicius G and Wang Y 2022 Improving reliability estimation for individual numeric predictions: a machine learning approach *INFORMS J. Comput.* **34** 503–21

[31] Abid A, Yuksekgonul M and Zou J 2022 Meaningfully debugging model mistakes using conceptual counterfactual explanations *Int. Conf. on Machine Learning* (PMLR) pp 66–88

[32] Vorontsov E *et al* 2024 A foundation model for clinical-grade computational pathology and rare cancers detection *Nat. Med.* **30** 2924–35

[33] Yuksekgonul M, Chandrasekaran V, Jones E, Gunasekar S, Naik R, Palangi H, Kamar E and Nushi B 2023 Attention satisfies: a constraint-satisfaction lens on factual errors of language models *12th Int. Conf. on Learning Representations*

[34] Hendrycks D and Gimpel K 2022 A baseline for detecting misclassified and out-of-distribution examples in neural networks *Int. Conf. on Learning Representations*

[35] Wang Y, Sun W, Jin J, Kong Z and Yue X 2023 WOOD: Wasserstein-based out-of-distribution detection *IEEE Trans. Pattern Anal. Mach. Intell.* **46** 944–56

[36] Chen R J, Wang J J, Williamson D F, Chen T Y, Lipkova J, Lu M Y, Sahai S and Mahmood F 2023 Algorithmic fairness in artificial intelligence for medicine and healthcare *Nat. Biomed. Eng.* **7** 719–42

[37] Liu J, Lin Z, Padhy S, Tran D, Bedrax Weiss T and Lakshminarayanan B 2020 Simple and principled uncertainty estimation with deterministic deep learning via distance awareness *Advances in Neural Information Processing Systems* vol 33 pp 7498–512

[38] Liu J Z, Padhy S, Ren J, Lin Z, Wen Y, Jerfel G, Nado Z, Snoek J, Tran D and Lakshminarayanan B 2023 A simple approach to improve single-model deep uncertainty via distance-awareness *J. Mach. Learn. Res.* **24** 1–63

[39] Ribeiro M T, Singh S and Guestrin C 2016 'Why should I trust you?' Explaining the predictions of any classifier *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1135–44

[40] Liang W, Tadesse G A, Ho D, Fei-Fei L, Zaharia M, Zhang C and Zou J 2022 Advances, challenges and opportunities in creating data for trustworthy AI *Nat. Mach. Intell.* **4** 669–77

[41] Jospin L V, Laga H, Boussaid F, Buntine W and Bennamoun M 2022 Hands-on bayesian neural networks—a tutorial for deep learning users *IEEE Comput. Intell. Mag.* **17** 29–48

[42] Zhang X and Mahadevan S 2020 Bayesian neural networks for flight trajectory prediction and safety assessment *Decis. Support Syst.* **131** 113246

[43] Mackay D J C 1992 *Bayesian Methods for Adaptive Models* (California Institute of Technology)

[44] Neal R M 2012 *Bayesian Learning for Neural Networks* vol 118 (Springer)

[45] Blundell C, Cornebise J, Kavukcuoglu K and Wierstra D 2015 Weight uncertainty in neural network *Int. Conf. on Machine Learning* (PMLR) pp 1613–22

[46] Graves A 2011 Practical variational inference for neural networks *Advances in Neural Information Processing Systems* vol 24

[47] Louizos C and Welling M 2016 Structured and efficient variational deep learning with matrix Gaussian posteriors *Int. Conf. on Machine Learning* (PMLR) pp 1708–16

[48] Gal Y and Ghahramani Z 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning *Int. Conf. on Machine Learning* (PMLR) pp 1050–9

[49] Gal Y and Ghahramani Z 2016 Bayesian convolutional neural networks with Bernoulli approximate variational inference
[50] Gal Y and Ghahramani Z 2016 A theoretically grounded application of dropout in recurrent neural networks *Advances in Neural Information Processing Systems* vol 29
[51] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems* vol 30
[52] Van Amersfoort J, Smith L, Jesson A, Key O and Gal Y 2021 On feature collapse and deep kernel learning for single forward pass uncertainty (arXiv:2102.11409)
[53] Van Amersfoort J, Smith L, Teh Y W and Gal Y 2020 Uncertainty estimation using a single deep deterministic neural network *Int. Conf. on Machine Learning* (PMLR) pp 9690–700
[54] Zhang X, Wang T, Yan C, Najdawi F, Zhou K, Ma Y, Cheung Y m and Malin B A 2024 Implementing trust in non-small cell lung cancer diagnosis with a conformalized uncertainty-aware AI framework in whole-slide images *medRxiv Preprint* https://doi.org/10.1101/2024.12.27.24319715 (posted online 30 December 2024)
[55] Huang Y, Song J, Wang Z, Zhao S, Chen H, Juefei-Xu F and Ma L 2025 Look before you leap: an exploratory study of uncertainty analysis for large language models *IEEE Trans. Softw. Eng.* **51** 413–29
[56] Chua M, Kim D, Choi J, Lee N G, Deshpande V, Schwab J, Lev M H, Gonzalez R G, Gee M S and Do S 2023 Tackling prediction uncertainty in machine learning for healthcare *Nat. Biomed. Eng.* **7** 711–8
[57] Banerji C R S, Chakraborti T, Harbron C and MacArthur B D 2023 Clinical AI tools must convey predictive uncertainty for each individual patient *Nat. Med.* **29** 2996–8
[58] Kompa B, Snoek J and Beam A L 2021 Second opinion needed: communicating uncertainty in medical machine learning *npj Digit. Med.* **4** 4
[59] Li J, Long X, Deng X, Jiang W, Zhou K, Jiang C and Zhang X 2024 A principled distance-aware uncertainty quantification approach for enhancing the reliability of physics-informed neural network *Reliab. Eng. Syst. Saf.* **245** 109963
[60] Thuy A and Benoit D F 2024 Explainability through uncertainty: trustworthy decision-making with neural networks *Eur. J. Oper. Res.* **317** 330–40
[61] Qu H, Foo L G, Li Y and Liu J 2023 Towards more reliable confidence estimation *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 13 152–13 169
[62] Sensoy M, Kaplan L and Kandemir M 2018 Evidential deep learning to quantify classification uncertainty *Advances in Neural Information Processing Systems* vol 31
[63] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B and Snoek J 2019 Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift *Advances in Neural Information Processing Systems* vol 32
[64] Linmans J, Elfwing S, van der Laak J and Litjens G 2023 Predictive uncertainty estimation for out-of- distribution detection in digital pathology *Med. Image Anal.* **83** 102655
[65] Zhu F, Zhang X-Y, Cheng Z and Liu C-L 2023 Revisiting confidence estimation: towards reliable failure prediction *IEEE Trans. Pattern Anal. Mach. Intell.* **46** 3370–87
[66] Corbière C, Thome N, Bar-Hen A, Cord M and Pérez P 2019 Addressing failure prediction by learning model confidence *Advances in Neural Information Processing Systems* vol 32
[67] Mucsányi B, Kirchhof M and Oh S J 2024 Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks (arXiv:2402.19460)
[68] Shafer G and Vovk V 2008 A tutorial on conformal prediction *J. Mach. Learn. Res.* **9** 371–421
[69] Angelopoulos A N, Bates S, Fannjiang C, Jordan M I and Zrnic T 2023 Prediction-powered inference *Science* **382** 669–74
[70] Balasubramanian V, Ho S-S and Vovk V 2014 *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications* (Newnes)
[71] Lei J, Rinaldo A and Wasserman L 2015 A conformal prediction approach to explore functional data *Ann. Math. Artif. Intell.* **74** 29–43
[72] Vovk V, Gammerman A and Shafer G 2005 *Algorithmic Learning in a Random World* vol 29 (Springer)
[73] Angelopoulos A N, Barber R F and Bates S2024 Theoretical foundations of conformal prediction (arXiv:2411.11824)
[74] Lei J, G'Sell M, Rinaldo A, Tibshirani R J and Wasserman L 2018 Distribution-free predictive inference for regression *J. Am. Stat. Assoc.* **113** 1094–111
[75] Romano Y, Sesia M and Candes E 2020 Classification with valid and adaptive coverage *Advances in Neural Information Processing Systems* vol 33 pp 3581–91
[76] Angelopoulos A N, Bates S, Jordan M and Malik J 2022 Uncertainty sets for image classifiers using conformal prediction *Int. Conf. on Learning Representations*
[77] Angelopoulos A N, Bates S, Fisch A, Lei L and Schuster T 2023 Conformal risk control *12th Int. Conf. on Learning Representations*
[78] Bates S, Angelopoulos A, Lei L, Malik J and Jordan M 2021 Distribution-free, risk-controlling prediction sets *J. ACM* **68** 1–34
[79] Angelopoulos A N, Bates S, Candès E J, Jordan M I and Lei L 2021 Learn then test: calibrating predictive algorithms to achieve risk control (arXiv:2110.01052)
[80] Barber R F, Candes E J, Ramdas A and Tibshirani R J 2023 Conformal prediction beyond exchangeability *Ann. Stat.* **51** 816–45
[81] Farinhas A, Zerva C, Ulmer D T and Martins A 2023 Non-exchangeable conformal risk control *12th Int. Conf. on Learning Representations*
[82] Vovk V 2012 Conditional validity of inductive conformal predictors *Asian Conf. on Machine Learning* (PMLR) pp 475–90
[83] Foygel Barber R, Candes E J, Ramdas A and Tibshirani R J 2021 The limits of distribution-free conditional predictive inference *Inf. Inference: J. IMA* **10** 455–82
[84] Gibbs I, Cherian J J and Candès E J 2023 Conformal prediction with conditional guarantees (arXiv:2305.12616)
[85] Jung C, Noarov G, Ramalingam R and Roth A 2023 Batch multivalid conformal prediction *Int. Conf. on Learning Representations (ICLR)*
[86] Blot V, Angelopoulos A N, Jordan M I and Brunel N J 2024 Automatically adaptive conformal risk control (arXiv:2406.17819)
[87] Ding T, Angelopoulos A, Bates S, Jordan M and Tibshirani R J 2024 Class-conditional conformal prediction with many classes *Advances in Neural Information Processing Systems* vol 36
[88] Javanmardi A, Stutz D and Hüllermeier E 2024 Conformalized credal set predictors (arXiv:2402.10723)
[89] Mohri C and Hashimoto T 2024 Language models with conformal factuality guarantees (arXiv:2402.10978)
[90] Chen T, Navrátil J, Iyengar V and Shanmugam K 2019 Confidence scoring using whitebox meta-models with linear classifier probes *22nd Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 1467–75
[91] Wolpert D H 1992 Stacked generalization *Neural Netw.* **5** 241–59
[92] Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A and Ueffing N 2004 Confidence

estimation for machine translation *Coling 2004: Proc. 20th Int. Conf. on Computational Linguistics* pp 315–21
[93] Corbiere C, Thome N, Saporta A, Vu T-H, Cord M and Perez P 2021 Confidence estimation via auxiliary models *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 6043–55
[94] Tsiligkaridis T 2021 Failure prediction by confidence estimation of uncertainty-aware Dirichlet networks *ICASSP 2021-2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 3525–9
[95] Tsiligkaridis T 2021 Information aware max-norm dirichlet networks for predictive uncertainty estimation *Neural Netw.* **135** 105–14
[96] Zhu F, Cheng Z, Zhang X-Y and Liu C-L 2022 Rethinking confidence calibration for failure prediction *European Conf. on Computer Vision* (Springer) pp 518–36
[97] Zhu F, Cheng Z, Zhang X-Y and Liu C-L 2023 Openmix: exploring outlier samples for misclassification detection *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* pp 12074–83
[98] Rabanser S, Thudi A, Hamidieh K, Dziedzic A and Papernot N 2022 Selective classification via neural network training dynamics (arXiv:2205.13532)
[99] Liu S, Ye H and Zou J 2025 Reducing hallucinations in large vision-language models via latent space steering *13th Int. Conf. on Learning Representations*
[100] Cheng Z, Zhu F, Zhang X-Y and Liu C-L 2024 Breaking the limits of reliable prediction via generated data *Int. J. Comput. Vis.* **133** 1–27
[101] Zhu F, Cheng Z, Zhang X-Y, Liu C-L and Zhang Z 2024 Rcl: reliable continual learning for unified failure detection *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* p 12 140–12 150
[102] Grabinski J, Gavrikov P, Keuper J and Keuper M 2022 Robust models are less over-confident *Advances in Neural Information Processing Systems* vol 35 p 39 059–39 075
[103] Zhang X-Y, Xie G-S, Li X, Mei T and Liu C-L 2023 A survey on learning to reject *Proc. IEEE* **111** 185–215
[104] Chow C K 1957 An optimum character recognition system using decision functions *IRE Trans. Electron. Comput.* **EC-6** 247–54
[105] Chow C 1970 On optimum recognition error and reject tradeoff *IEEE Trans. Inf. Theory* **16** 41–46
[106] Elkan C 2001 The foundations of cost-sensitive learning *Int. Joint Conf. on Artificial Intelligence* vol 17 (Lawrence Erlbaum Associates Ltd) pp 973–8
[107] Mao A, Mohri M and Zhong Y 2024 Theoretically grounded loss functions and algorithms for score-based multi-class abstention *Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 4753–61
[108] Cordella L P, De Stefano C, Tortorella F and Vento M 1995 A method for improving classification reliability of multilayer perceptrons *IEEE Trans. Neural Netw.* **6** 1140–7
[109] De Stefano C, Sansone C and Vento M 2000 To reject or not to reject: that is the question-an answer in case of neural classifiers *IEEE Trans. Syst. Man Cybern. C* **30** 84–94
[110] Charoenphakdee N, Cui Z, Zhang Y and Sugiyama M 2021 Classification with rejection based on cost-sensitive classification *Int. Conf. on Machine Learning* (PMLR) pp 1507–17
[111] Nguyen V-L and Hullermeier E 2020 Reliable multilabel classification: prediction with partial abstention *Proc. AAAI Conf. on Artificial Intelligence* vol 34 pp 5264–71
[112] Kalai A and Kanade V 2021 Towards optimally abstaining from prediction with OOD test examples *Advances in Neural Information Processing Systems* vol 34 p 12 774–12 785
[113] Geifman Y and El-Yaniv R 2017 Selective classification for deep neural networks *Advances in Neural Information Processing Systems* vol 30
[114] Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G and Mohd-Yusof J 2019 Combating label noise in deep learning using abstention *Int. Conf. on Machine Learning* (PMLR) pp 6234–43
[115] Thulasidasan S 2020 *Deep Learning With Abstention: Algorithms for Robust Training and Predictive Uncertainty* (University of Washington)
[116] Geifman Y and El-Yaniv R 2019 SelectiveNet: a deep neural network with an integrated reject option *Int. Conf. on Machine Learning* (PMLR) pp 2151–9
[117] Guo Y, Jiao F, Shen Z, Nie L and Kankanhalli M 2024 UNK-VQA: a dataset and a probe into the abstention ability of multi-modal large models *IEEE Trans. on Pattern Analysis and Machine Intelligence*
[118] Ehlers R 2017 Formal verification of piece-wise linear feed-forward neural networks *Automated Technology for Verification and Analysis: 15th Int. Symp., ATVA 2017 (Pune, India, 3–6 October 2017), Proc. 15* (Springer) pp 269–86
[119] Katz G *et al* 2019 The marabou framework for verification and analysis of deep neural networks *Computer Aided Verification: 31st Int. Conf., CAV 2019, (New York City, NY, USA, 15–18 July 2019), Proc., Part I 31* pp 443–52
[120] Huang P, Wu H, Yang Y, Daukantas I, Wu M, Zhang Y and Barrett C 2024 Towards efficient verification of quantized neural networks *Proc. AAAI Conf. on Artificial Intelligence* vol 38 pp 21152–60
[121] Albarghouthi A *et al* 2021 Introduction to neural network verification *Found. Trends® Program. Lang.* **7** 1–157
[122] Pulina L and Tacchella A 2010 An abstraction-refinement approach to verification of artificial neural networks *Computer Aided Verification: 22nd Int. Conf., CAV 2010, (Edinburgh, UK, 15–19 July 2010) Proc. 22* (Springer) pp 243–57
[123] Huang X, Kwiatkowska M, Wang S and Wu M 2017 Safety verification of deep neural networks *Computer Aided Verification: 29th Int. Conf., CAV 2017, (Heidelberg, Germany, 24–28 July 2017), Proc., Part I 30* (Springer) pp 3–29
[124] Sun X, Khedr H and Shoukry Y 2019 Formal verification of neural network controlled autonomous systems *Proc. 22nd ACM Int. Conf. on Hybrid Systems: Computation and Control* pp 147–56
[125] Venzke A and Chatzivasileiadis S 2020 Verification of neural network behaviour formal guarantees for power system applications *IEEE Trans. Smart Grid* **12** 383–97
[126] Katz G, Barrett C, Dill D L, Julian K and Kochenderfer M J 2017 Reluplex: an efficient SMT solver for verifying deep neural networks *Computer Aided Verification: 29th Int. Conf., CAV 2017, (Heidelberg, Germany, 24–28 July 2017), Proc., Part I 3* (Springer) pp 97–117
[127] Wang S, Pei K, Whitehouse J, Yang J and Jana S 2018 Formal security analysis of neural networks using symbolic intervals *27th USENIX Security Symp. (USENIX Security 18)* pp 1599–614
[128] Santa Cruz U and Shoukry Y 2022 NNLander-VeriF: a neural network formal verification framework for vision-based autonomous aircraft landing *NASA Formal Methods Symp.* (Springer) pp 213–30
[129] Wang S, Pei K, Whitehouse J, Yang J and Jana S 2018 Efficient formal safety analysis of neural networks *Advances in Neural Information Processing Systems* vol 31
[130] Urban C and Miné A 2021 A review of formal methods applied to machine learning (arXiv:2104.02466)

[131] Von Rueden L *et al* 2021 Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems *IEEE Trans. Knowl. Data Eng.* **35** 614–33
[132] Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S and Yang L 2021 Physics-informed machine learning *Nat. Rev. Phys.* **3** 422–40
[133] Raissi M, Perdikaris P and Karniadakis G E 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations *J. Comput. Phys.* **378** 686–707
[134] Xu Y, Kohtz S, Boakye J, Gardoni P and Wang P 2023 Physics-informed machine learning for reliability and systems safety applications: state of the art and challenges *Reliab. Eng. Syst. Saf.* **230** 108900
[135] Zhou T, Zhang X, Droguett E L and Mosleh A 2023 A generic physics-informed neural network-based framework for reliability assessment of multi-state systems *Reliab. Eng. Syst. Saf.* **229** 108835
[136] Cuomo S, Di Cola V S, Giampaolo F, Rozza G, Raissi M and Piccialli F 2022 Scientific machine learning through physics–informed neural networks: where we are and what's next *J. Sci. Comput.* **92** 88
[137] Thelen A, Zhang X, Fink O, Lu Y, Ghosh S, Youn B D, Todd M D, Mahadevan S, Hu C and Hu Z 2022 A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies *Struct. Multidiscip. Optim.* **65** 354
[138] Yucesan Y A and Viana F A 2021 Hybrid physics-informed neural networks for main bearing fatigue prognosis with visual grease inspection *Comput. Ind.* **125** 103386
[139] Dong P, Wang X L, Bose I, Ng K K, Zhang X and Zhang X 2024 Causally-aware spatio-temporal multi-graph convolution network for accurate and reliable traffic prediction (arXiv:2408.13293)
[140] Kyono T, Zhang Y and van der Schaar M 2020 Castle: regularization via auxiliary causal graph discovery *Advances in Neural Information Processing Systems* vol 33 pp 1501–12
[141] Teshima T and Sugiyama M 2021 Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation *Uncertainty in Artificial Intelligence* (PMLR) pp 86–96
[142] Kancheti S S, Reddy A G, Balasubramanian V N and Sharma A 2022 Matching learned causal effects of neural networks with domain priors *Int. Conf. on Machine Learning* (PMLR) p 10 676–10 696
[143] Wen B, Cao Y, Yang F, Subbalakshmi K and Chandramouli R 2022 Causal-TGAN: modeling tabular data using causally-aware GAN *ICLR Workshop on Deep Generative Models for Highly Structured Data*
[144] Cui P and Athey S 2022 Stable learning establishes some common ground between causal inference and machine learning *Nat. Mach. Intell.* **4** 110–5
[145] Zhang X, Cui P, Xu R, Zhou L, He Y and Shen Z 2021 Deep stable learning for out-of-distribution generalization *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* pp 5372–82
[146] Zhai P, Yang Y and Zhang C 2023 Causality-based CTR prediction using graph neural networks *Inf. Process. Manage.* **60** 103137
[147] Berrevoets J, Kacprzyk K, Qian Z and van der Schaar M 2023 Causal deep learning (arXiv:2303.02186)
[148] de Beaulieu M H, Jha M S, Garnier H and Cerbah F 2024 Remaining useful life prediction based on physics-informed data augmentation *Reliab. Eng. Syst. Saf.* **252** 110451
[149] Kapusuzoglu B and Mahadevan S 2020 Physics-informed and hybrid machine learning in additive manufacturing: application to fused filament fabrication *JOM* **72** 4695–705
[150] Pei K, Cao Y, Yang J and Jana S 2017 Deepxplore: automated whitebox testing of deep learning systems *Proc. 26th Symp. on Operating Systems Principles* pp 1–18
[151] Dwivedi R *et al* 2023 Explainable AI (XAI): core ideas, techniques and solutions *ACM Comput. Surv.* **55** 1–33
[152] Gunning D and Aha D 2019 Darpa's explainable artificial intelligence (XAI) program *AI Mag.* **40** 44–58
[153] Sundararajan M and Najmi A 2020 The many shapley values for model explanation *Int. Conf. on Machine Learning* (PMLR) pp 9269–78
[154] Zintgraf L M, Cohen T S, Adel T and Welling M 2017 Visualizing deep neural network decisions: prediction difference analysis *Int. Conf. on Learning Representations*
[155] Zhang X, Chan F T and Mahadevan S 2022 Explainable machine learning in image classification models: an uncertainty quantification perspective *Knowl.-Based Syst.* **243** 108418
[156] Kuznietsov A, Gyevnar B, Wang C, Peters S and Albrecht S V 2024 Explainable AI for safe and trustworthy autonomous driving: a systematic review *IEEE Trans. Intell. Transp. Syst.* **25** 19342–64
[157] Bau D, Zhu J-Y, Strobelt H, Lapedriza A, Zhou B and Torralba A 2020 Understanding the role of individual units in a deep neural network *Proc. Natl Acad. Sci.* **117** 30071–8
[158] Samek W, Montavon G, Vedaldi A, Hansen L K and Müller K-R 2019 *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* vol 11700 (Springer Nature)
[159] Zintgraf L M, Cohen T S, Adel T and Welling M 2022 Visualizing deep neural network decisions: prediction difference analysis *Int. Conf. on Learning Representations*
[160] Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15
[161] Sinha R, Schmerling E and Pavone M 2023 Closing the loop on runtime monitors with fallback-safe MPC *2023 62nd IEEE Conf. on Decision and Control (CDC)* (IEEE) pp 6533–40
[162] Lewis P *et al* 2020 Retrieval-augmented generation for knowledge-intensive NLP tasks *Advances in Neural Information Processing Systems* vol 33 pp 9459–74

**Xiaoge Zhang** is an Assistant Professor in the Department of Industrial and Systems Engineering (ISE) at The Hong Kong Polytechnic University. He received his PhD in Systems Engineering and Operations Research at Vanderbilt University, Nashville, Tennessee, United States in 2019. He has won multiple awards, including the Peter G. Hoadley Best Paper Award, the Chinese Government Award for Outstanding Self-Financed Students Studying Abroad, the Bravo Zulu Award, Pao Chung Chen Fellowship, to name a few. He has published more than 80 papers in leading academic journals, such as *Nature Communications*, *IEEE Transactions on Reliability*, *IEEE Transactions on Artificial Intelligence, IEEE Transactions on Automation Science and Engineering*, *Reliability Engineering & Systems Safety*, *Risk Analysis*, *IEEE Transactions on Industrial Informatics*, *Decision Support Systems*, *IEEE Transactions on Cybernetics, and Annals of Operations Research*, among others. His research has gathered widespread attention from the academic community (3000+ citations, h-index 34 according to Google Scholar). He is a member of IEEE and IISE. His research interests center on safety,

reliability and trustworthiness assurance of AI-powered intelligent systems and their applications to high-stakes decision settings.

**Tao Wang** received his Bachelor's degree in Automation and Master's degree in Control Engineering from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2020 and 2023, respectively. He is currently pursuing a PhD degree in the Department of Industrial and Systems Engineering at The Hong Kong Polytechnic University, Hong Kong, China. His research interests include conformal prediction, uncertainty quantification, and industrial AI applications.

**Lei Ma** received a BE degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, and the ME and PhD degrees from The University of Tokyo, Tokyo, Japan, in 2011 and 2014, respectively. He is currently an Associate Professor at The University of Tokyo and the University of Alberta, Edmonton, AB, Canada. He was honorably selected as Canada CIFAR AI Chair and a fellow with Alberta Machine Intelligence Institute (Amii), Edmonton. His research interests include the interdisciplinary fields of software engineering (SE) and trustworthy artificial intelligence, with a special focus on the quality, reliability, safety, and security aspects of AI systems.

**Sankaran Mahadevan** has thirty-five years of research and teaching experience in uncertainty quantification, risk and reliability analysis, machine learning, system health diagnosis and prognosis, and optimization under uncertainty. His research has been extensively funded by NSF, NASA, FAA, DOE, DOD, DOT, NIST, General Motors, Chrysler, Union Pacific, American Railroad Association, and Sandia, Idaho, Los Alamos and Oak Ridge National Laboratories. His research contributions are documented in more than 450 publications, including two textbooks on reliability methods and 200 journal papers. He has directed 40 PhD dissertations and 24 MS theses, and has taught many industry short courses on reliability and risk analysis methods. During the past decade, he has been at the forefront of academic research on digital twin methodologies for air and marine transportation vehicles, buildings, additive manufacturing, and power grid networks. He is currently the President of the ASCE Engineering Mechanics Institute, Managing Editor of the ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, and a winner of the Senior Distinguished Research Award from the International Association of Structural Safety and Reliability. He is a Fellow of the American Institute of Aeronautics and Astronautics (AIAA), Engineering Mechanics Institute (ASCE), and the Prognostics & Health Management Society. His awards include the NASA Next Generation Design Tools award (NASA), the SAE Distinguished Probabilistic Methods Educator Award, and best paper awards in the MORS Journal and the SDM and IMAC conferences. Professor Mahadevan obtained his B.S. from the Indian Institute of Technology, Kanpur, M.S. from Rensselaer Polytechnic Institute, Troy, NY, and PhD from Georgia Institute of Technology, Atlanta, GA.