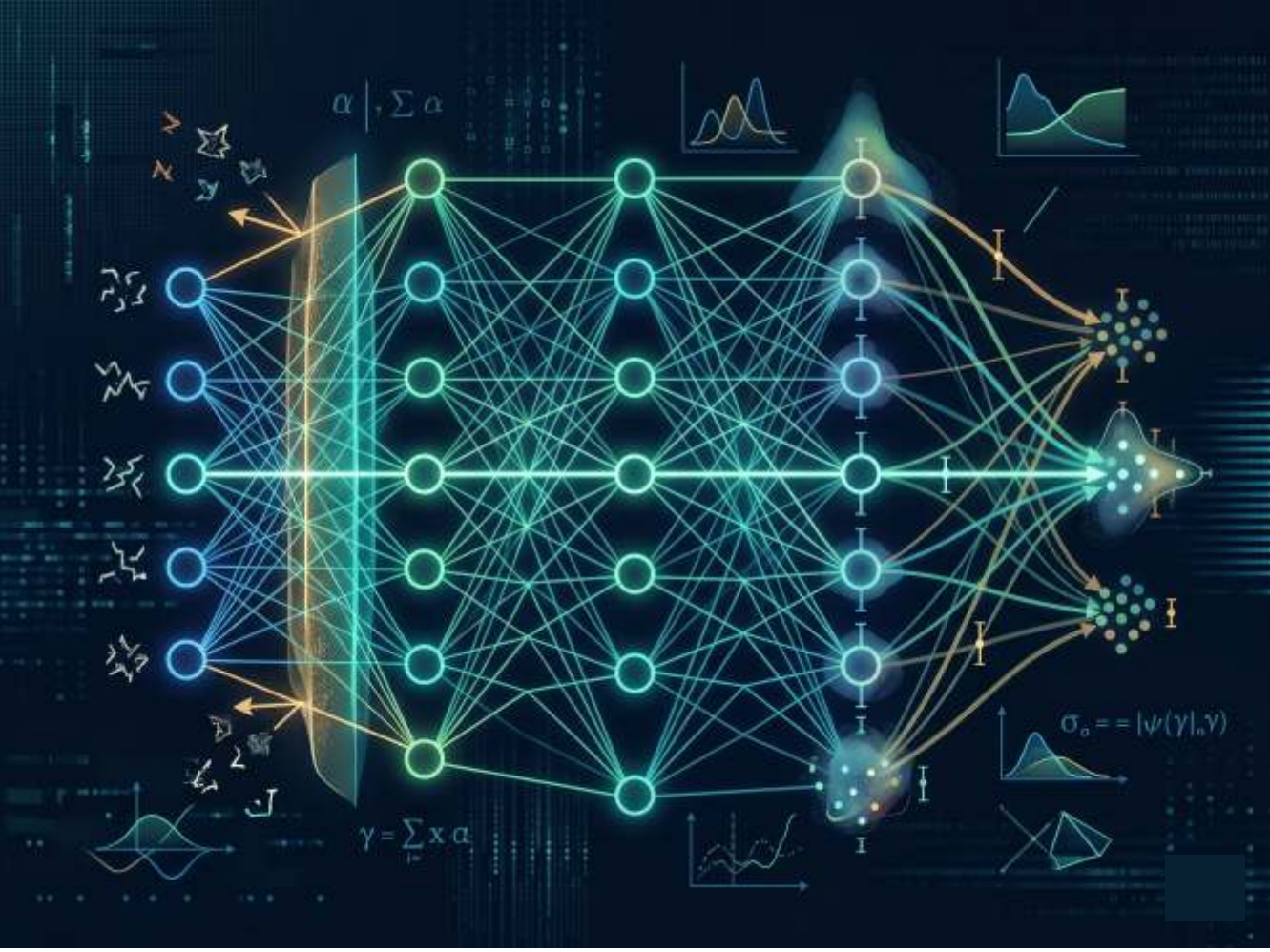

# Trustworthy Deep Learning

## Robustness, Uncertainty Quantification, and Adversarial Resilience

Karn Singh
Puneet Garg

DeepScience

# Trustworthy Deep Learning: Robustness, Uncertainty Quantification, and Adversarial Resilience

**Karn Singh**

St. Andrews Institute of Technology and Management, Gurugram-122506, Haryana, India

**Puneet Garg**

KIET (Deemed to be University), Delhi NCR, Ghaziabad-201206, Uttar Pradesh, India

DeepScience

*Published, marketed, and distributed by:*

Deep Science Publishing, 2026
USA | UK | India | Turkey
Reg. No. MH-33-0658412
www.deepscienceresearch.com
editor@deepscienceresearch.com
WhatsApp: +91 7977171947

ISBN: 978-93-7185-764-2

E-ISBN: 978-93-7185-510-5

https://doi.org/10.70593/978-93-7185-510-5



**Citation:** Singh, K., & Garg, P. (2026). *Trustworthy Deep Learning: Robustness, Uncertainty Quantification, and Adversarial Resilience*. Deep Science Publishing. https://doi.org/10.70593/978-93-7185-510-5

# Preface

Deep learning has become one of the most vital infrastructures under study and entirely a subject of research not at a slow rate that will allow us to comprehend its failures. We apply neural networks to hospitals, financial markets, autonomous systems, and so on, but the culture of the field has traditionally been that benchmark accuracy is the most valued criterion. This book is a corrective. Credible AIs are based on three pillars that are interdependent. Robustness poses whether a model will continue to handle purely outside the regime of its training distribution. Uncertainty quantification inquires a question on whether a model knows which it does not know. Adversarial resilience. The question of adversarial resilience is whether a model can be resilient to its intended manipulation. Each of the challenges is severe in itself; the combination of all of them determines the difference between a model that is functional in the laboratory and one that is worth trusted in the real world.

The book is structured in this regard, having three parts, each of them representing the pillar, and a final section, the evaluation frameworks and open problems. The chapters are concluded by exercises as well as annotated literature pointers. The companion repository contains the code of the core examples. It presupposes the knowledge of the basics of deep learning among the readers. It does not need any specific experience in Bayesian statistics or adversarial machine learning. Graduate learners, those who work in the high-stakes fields, and researchers who seek to make cross-cutting in the three areas will all have something appropriate to them.

We are honest in our writing of what cannot and can be done by the current method. Strategies that seemed to be good have again and again been defeated by an adaptive attack; distributions that seemed to be estimated have been wrecked by distribution shift. Where evidence is great, we say so. Where it is amalgamated, we say that too. Instead of an optimistic map of the frontier an accurate map of the frontier is of better service to the reader. Deep learning has provided us with unprecedented abilities. There is work to do in order to get those capabilities dependable, truthful about their capabilities and inaccessible to abuse. Hopefully, this book is one of these steps.

Karn Singh
Puneet Garg

# Table of Contents

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

# Chapter 1: Trustworthy Deep Learning: Motivation, Challenges, and Applications

## Abstract

Deep learning has had impressive empirical achievements in many areas, but its empirical use in high-stakes industries has shown fundamental weaknesses in terms of reliability, safety and accountability. In this chapter, the author presents the conceptual architecture of trustworthy deep learning that addresses what motivates the study on the topic, the technical and sociotechnical difficulties that cannot be ignored, and the application areas that require not only trustworthiness but also its existence. Using the latest breakthroughs in the field of robustness theory, quantity of uncertainty, adversarial machine learning, learning with fairness, and interpretability, the chapter poses these issues in a single context that focuses on the complementary relation between the technical rigor and the ethical responsibility. Critical review of the emerging regulation stage and its implications to model development is done in the chapter also, and indicators of open problems which demarcate the frontier of the field are also indicated.

**Keywords**: AI, the robustness of deep learning, uncertainty quantification, adversarial attacks, distributional shift, interpretability, the AI safety and responsible AI.

## 1. Deep Learning: the Trustworthiness Imperative.

Deep learning is one of the most fateful technology shifts of the twenty-first century. It has been observed that since the groundbreaking experiment by Krizhevsky, Sutskever, and Hinton in 2012 showing that deep convolutional neural networks could significantly outperform classical computer vision approaches with respect to large-scale image recognition tasks, the field has seen an unparalleled spread over deep learning architectures, training algorithms and tasks to apply (LeCun, Bengio, and Hinton, 2015). Deep neural networks nowadays are the foundation of language translation, protein structure prediction, autonomous navigation, medical diagnosis, financial risk

assessment, and an ever-expanding number of other consequential decision-making systems. These models are so capable of prediction, however, that enormous financial investments have been created by solely their sheer predictive power, as embodied in their effort to extract complex statistical regularities out of large body of data.

However, with this rise, a corresponding and growing corpus of literature has established the fragility, opaqueness, and the possible harm that defines existing systems of deep learning. The groundbreaking finding of Szegedy et al. (2013) that neural networks of state-of-the-art accuracy would misclassify input images with high confidence in response to imperceptibly small and well-designed perturbations to the input images, the so-called adversarial examples, provoked a surge of research activities into the underlying structural weaknesses of neural networks. Later investigations demonstrated that these weaknesses were not just another casualty of curiosity, but signs of more fundamental epistemic drawbacks: deep networks learned using common empirical risk minimization processes are ill-calibrated, subject to distribution shift, sensitive to spurious correlations, and unable to reliably estimate their own uncertainty. Such restrictions are not harmless in low stakes contexts; they may be disastrous in any system, when false forecasts have permanent effects on a physical, economic, or social system.

The idea of reliability in AI and deep learning has thus become a converging research agenda that does not reflect one of the technical issues. The notion of trustworthiness includes, but is not limited to, the concepts of robustness (the ability to operate under the presence of distribution shift and adversarial perturbation), calibration and uncertainty quantification (ability to make reliable confidence estimates and express ignorance when it is relevant), fairness (equitable behavior depending on demographic groups and protected attributes), interpretability and explainability (transparency in the inference processes generating predictions), privacy (sensitive information in training data and model parameters), and accountability (have clear responsibility on the behavior of the model). These dimensions are not that independent, as we shall illustrate in this book, as progress in one field often has consequences of one kind or another in others, either positive or negative.

The chapter plays the role of the conceptual basis of the rest of the book. It starts by putting the plausible deep learning agenda in this larger intellectual and social situation and analyzes the colliding forces of scholarly investigation, regulation, and in-world performance failures that have elevated this subject a subject of focus in artificial intelligence research in the 2020s. It then gives a systematic taxonomy of the technical issues involved that should be resolved, disclosing the means in which the existing systems of deep learning fall short of the standard of trustworthiness. The chapter continues with an overview of the most notable areas of application where reliable deep learning is absolutely mandatory, providing specific instances on how the abstract

technical issues can be generalized into practical threats and advantages. It ends with the summary of the organizational structure of the book, which puts the further chapters in the context of the larger picture created here.

## 2. Incentives of Reliable Deep Learning.

### 2.1 The Gap of Deployment and Real-Life Failures.

The motivation behind credible deep learning research is driven and initially and most tangibly by a collection of failures of deployed machine learning systems that have been documented in the past. In medicine, researchers proved that convolutional neural networks to detect diabetic retinopathy based on fundus image every represented significant image deterioration in performance when applied to images of demographic groups and cameras (Finlayson et al., 2019). Adversarial patches Painted on stoplights On in self-driving, neural network perception systems were found to misclassify adversarial patches, a physical object that could be attacked to offer happenings related to the security of significant infrastructure (Eykholt et al., 2018). The error rates of commercial facial recognition systems sold by large technology providers were many times greater in case of dark-skinned women compared to light-skinned males, which could be explained by the imbalance in training data (Buolamwini and Gebru, 2018). The algorithms of recidivism prediction applied in the criminal justice system of the United States were reported to have had racially imbalanced false positive results which incriminated Black defendants (Angwin et al., 2016). These are not incidental failures, as they constitute systematic demonstrations that there is a wide and significant gap between the controlled benchmark performance and real-world trust worthiness.

The theme in these failures is that the basic assumption of standard machine learning is that the distribution of training data is sufficient to reflect on the distribution of the deployment conditions. This is the so called i.i.d. ( independent and identically distributed ) assumption and this is frequently realized not to be true in practice. Healthcare data are gathered on particular hospital systems, imaging equipment, and patient groups and extrapolation to different clinical contexts causes covariate shift that can adversely affect even the most reliable training-time models and render them unreliable in practice. Autonomous systems are subject to weather conditions, road geometries as well as sensor degradations which are not covered by training distributions. The biases, toxicity, and misinformation encoded by language models which are trained with internet text and when given queries that are not in accordance with the training conditions will introduce confidently incorrect or harmful behaviors. The implementation discrepancy or the difference between the benchmark accuracy and

the real-world reliability is therefore the rationale behind the reliable deep learning program.

## 2.2 The Regulatory and Governance Imperative.

The trustworthiness agenda has been expedited forcefully by the factors other than the empirical history of failure; the existence of a global regulatory arena, which is increasingly imposing provable AI safety and responsibility. The most widespread process of attempting to regulate AI systems on a risk-based framework is the Artificial Intelligence Act of the European Union (EU AI Act) that came into force in 2024, and aspects of which continue to be implemented through 2026 and beyond.

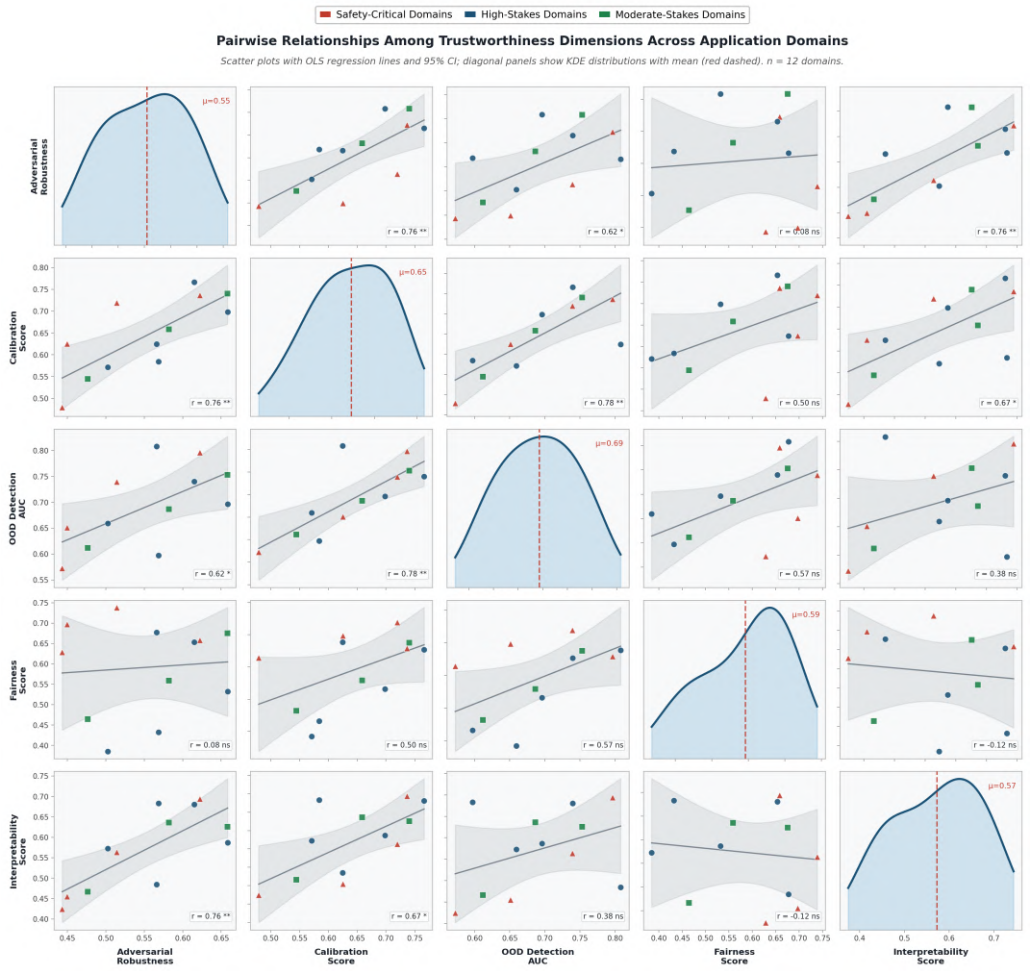


Fig. 1 Pairwise Scatter Matrix with KDE Marginals

Fig.1 shows A 5×5 symmetric scatter plot matrix showing all pairwise relationships among the five trustworthiness dimensions (Adversarial Robustness, Calibration Score, OOD Detection AUC, Fairness Score, and Interpretability Score) across 12 application domains. Diagonal panels display KDE density curves with mean (red dashed line) for each dimension. Off-diagonal cells show domain-coloured scatter points (by risk category), OLS regression lines, 95% confidence bands, and Pearson *r* with significance stars. This figure reveals multicollinear structure among dimensions and highlights which combinations exhibit the strongest tradeoffs.

The Act contains categories of AI systems prohibited, high-risk, limited-risk and minimal-risk systems, and creates strict requirements on developers and deployers of high-risk AI systems that affect sectors such as medical devices, critical infrastructure, law enforcement, migration, education, and employment. Such requirements are conformity assessment, risk management system, data governance as well as transparency, human oversight, accuracy, robustness and cybersecurity. Importantly, the EU AI Act does not only demand substantial accuracy; it demands that it is documented that systems are accurate and reliable in conditions of misuse and distributional variation that are reasonably predictable-demands that are directly allied to the technical agenda of responsible deep learning.

Similar regulatory processes in the United States, such as the AI Risk Management Framework (AI RMF 1.0) by the National Institute of Standards and Technology (NIST) (publicated in 2023), the Executive Order on Safe, Secure, and Trustworthy AI (internet) by the Biden administration (October 2023), and the following AI action plan (2025) by the Trump administration, indicate that both parties have agreed that the AI systems deployed in consequential areas need systematic risk management. The financial regulatory authorities (the Federal Reserve, the Office of the Comptroller of the Currency, the Consumer Financial Protection Bureau) have also offered regulatory guidance on model risk management that has more and more focused on machine learning and deep learning based models in credit underwriting, fraud detection and market-making. Other requirements of financial institutions implementing AI systems are put on by the Principles on Operational Resilience issued by the Basel Committee on Banking Supervision, and guidance on how to manage third-party risks. This can be seen with similar regulatory pressure in the healthcare sector (FDA guidance on AI/ML-based Software as a Medical Device), aviation sector (EASA AI roadmap), and the defense sector (NATO AI principles, US DoD Directive 3000.09 on autonomous weapons). The overlapping of regulatory pressure between jurisdictions and sectors has made trustworthy deep learning a research demand, rather than a commercial or legal one.

### 2.3 Epistemological Motivations and Scientific Motivations.

In addition to the practical requirements of safety in deploying and regulatory issues, the reliable deep learning agenda is driven by deep scientific inquiries of how learning, generalization, and representation of knowledge in neural networks take place. Classical statistical learning theory has been formulated in the context of Vapnik-Chervonenkis (VC) theory and Rademacher complexity that can give information about the deterministic origin of the gap between the empirical and population risk- these bounds, however, are vacuous in the over-parameterized deep networks that are currently performing well in practice (Zhang et al., 2017). The finding that deep networks are capable of perfectly memorizing randomly labeled training data without depreciating on structured data queries intuitive understanding of the bias-variance tradeoff as well as suggests that unstructured generalization theory cannot explain or predict the behavior of deep learning systems. This theoretical lacuna is the reason why a program of research exists to study the geometry of loss landscapes, implicit regularization and stochastic gradient descent, the connection between overparameterizing and robustness, and structural properties of deep representations that can or cannot support generalization.

This problem of deep learning in quantifying the uncertainty is also similarly driven by underlying epistemological factors. When the model predictions are confident, it is the well-calibrated model that ought to exist, and when there is uncertainty, the well-calibrated model ought to exist. It is also known that standard deep networks optimized using maximum likelihood estimation are overconfident: they place mass high probability on individual predictions even when the prediction is made on an input that is significantly outside the training distribution, a phenomenon commonly called epistemic overconfidence or uncertainty miscalibration (Guo et al., 2017). Not only is this calibration failure a technical inconvenience, but the consequences of the failure are far-reaching when it comes to the downstream use of model outputs in decision-making. A system that medical diagnoses with 97 percent probability on the incorrect classification does not give any indication to a clinician that the judgment of the model ought to be doubted, but a system that states that it is highly uncertain by 60 percent encourages the human oversight. The scientific objective of quantifying uncertainty in deep learning can therefore be expressed to be quantifying the Bayesian epistemological virtues of neural networks to distinguish what is known and what is not known and conveying that knowledge to downstream decision-makers.

## 3. Technical Problems in attaining Trustworthiness.

### 3.1 Adversarial Vulnerability and Robustness:

This is the field of research that deals with the issues of machine information systems ethics.

Adversarial vulnerability up to imperceptible targeted perturbations that lead to catastrophic misclassification is one phenomenon that has been most difficult in technical aspects to learn and practically important to overcome to achieve reliable deep learning. Based on the description of adversarial vulnerability by Goodfellow, Shlens, and Szegedy (2015) via the linearity hypothesis and the Abstract of the Fast Gradient Sign Method (FGSM), there have been now ample adversarial attacks, including white-box attacks where the full model is available (C&W, PGD/Madry attacks), black-box attacks identified as involving transferability or query-based optimization (ZOO, HopSkipJump), and attacks more advanced toward large language models, such as prompt injection and jailbreaking and indirect prompt injection through tool Projected gradient descent (PGD) adversarial attack of Madry et al. (2018) introduced the conventional threat model used in robustness analysis: a threat that can be limited to perturbations whose norm is less than p attempts to optimize model loss. This formalization made possible the creation of the adversarial training, the strategy of training data enrichment with adversarial perturbed examples, as the most common empirical defense mechanism.

Although it has been shown that much has changed, adversarial robustness is inherently challenging due to a number of factors that are mutually reinforcing. On the one hand, clean accuracy and adversarial robustness are intrinsically at odds: models that are participants in the adversarial robustness threat model often perform poorly on clean data or other perturbation modes, the so-called robustness-accuracy tradeoff (Raghunathan et al., 2020; Tsipras et al., 2019). Second, Adaptive attacks make the analysis of a robustness difficult: many defenses claimed to be robust when manipulating the attack under standard attack protocols are defeated when the attacker adapts to the particular defense mechanism, which highlights the importance of the rigorous, adaptive evaluation methodology (Tramer et al., 2020). Third, certified defenses - that can give mathematical assurances that no perturbation in some specified norm ball can result in misclassification - are computationally costly and certifiable radii tend to be significantly smaller than perturbation budgets used in practice. Approaches such as randomized smoothing (Cohen et al., 2019) and interval bound propagation are potentially valuable steps towards useful certified robustness, although extending the methods to large networks and high-dimensional inputs and ensuring that the certified radii have any useful scale remains a research challenge. The advent of foundation models and large language models also introduces new adversarial vulnerability dimensions, like membership inference attack, model extraction, and prompt-based adversarial manipulation which go far beyond the image-domain threat models on which the initial work on robustness has been based.

### 3.2 Distributional Shift and Generalization out of distribution.

Distributional shift Distributional shift is the discrepancy between the distribution of the data that is seen during training and deployment, and is arguably the most widespread basis of unreliability in deployed deep learning systems. Distributional shift may be brought by covariate shift (the marginal distribution of inputs is changed, but the conditional distribution of outputs given inputs is unchanged), label shift (the marginal distribution of outputs is changed), concept drift (the conditional distribution of outputs given inputs changes with time), or domain shift (a mix of the above due to a change in the environment, populations, or the tools of measurement used to collect data). Both forms of changes present different reliability issues with the model, and they need different technical resolutions. Ordinary empirical risk minimization aims to minimize risk that is expected given the training distribution and gives no assurances about the performance given shifted distributions, in fact, in the case of a severe shift, a model can be arbitrarily worse than a naive baseline.

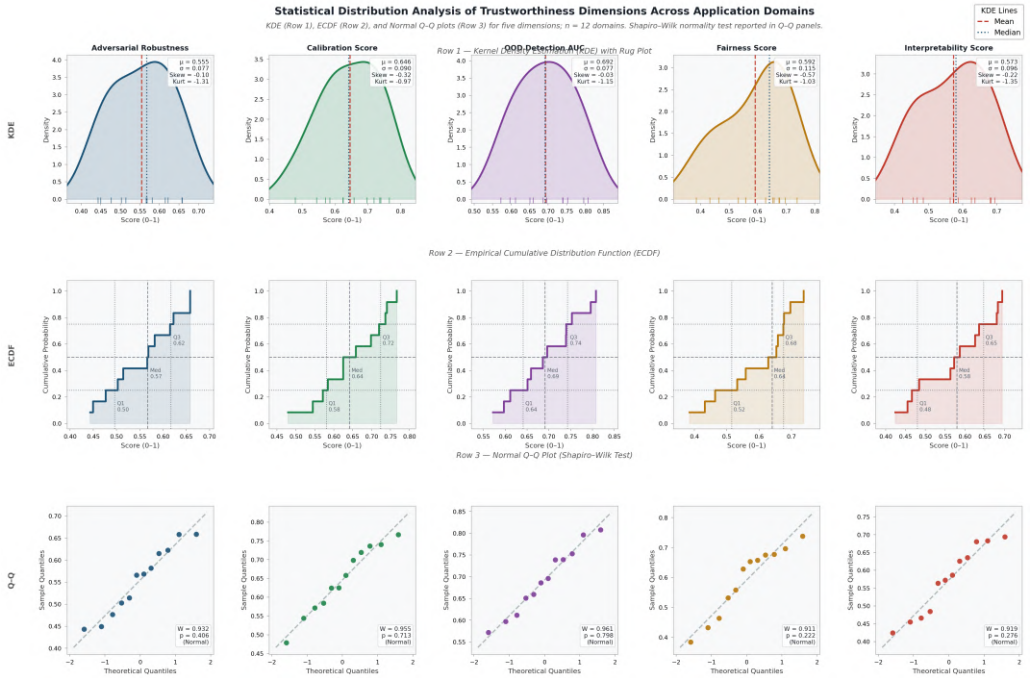


Fig. 2 Statistical Distribution Analysis (KDE / ECDF / Q–Q)

Above fig explains A 3-row × 5-column panel figure providing a complete distributional characterisation of each trustworthiness dimension. Row 1 shows KDE plots with rug marks, descriptive statistics (mean, SD, skewness, kurtosis), and mean/median reference lines. Row 2 shows Empirical Cumulative Distribution Functions (ECDFs) with

Q1/Median/Q3 annotations. Row 3 presents Normal Q–Q plots with Shapiro–Wilk test results, assessing distributional normality—critical information for selecting appropriate inferential tests.

Domain generalization Domain generalization aims at designing training algorithms such that they generate models that tolerate a given type of distribution shift in the absence of training time access to target domain data. Invariant risk minimization (Arjovsky et al., 2019), distributionally robust optimization (DRO), data augmentation strategies to explicitly extend the training distribution, and self-supervised and contrastive pre-training are some of the important strategies. Although it has come far, domain generalization is still an unresolved problem: algorithms that have high scores on Benchmark datasets like DomainBed (Gulrajani & Lopez-Paz, 2020) do not always perform better than simple empirical risk minimization with the correct regularization across all classes of distribution shifts, which adds to the overall challenge of learning representations that are useful across distribution classes unspecified in advance. Test-time adaptation techniques: learning model parameters or predictions with unlabeled test data constitute a growing competing approach to trainingtime domain generalization, with excellent corruption benchmark results of ImageNet-C (Hendrycks and Dietterich, 2019) and new stability, overfitting to test noise, and security concerns.

### 3.3 Quantification and Calibration of uncertainty.

The uncertainty quantification (UQ) problem in deep learning has two conceptually distinct but practically confounded issues: calibration (the accuracy rates on empirical accuracy matches model much more than they match empirical accuracy) and the problem of epistemic uncertainty estimation (is the uncertainty due to a finite data set or model imperfections). Classical Bayesian inference offers a rigorous theory of both: a Bayesian model will keep a posterior distribution over the parameters, and the natural offspring of this will be the predictive uncertainty, the dispersion of the predictive distribution. Nevertheless, deep networks can have millions to billions of parameters and thus cannot be the precise target of bayesian inference computationally, requiring approximations. Markov Chain Monte Carlo (MCMC) methods such as Hamilton Monte Carlo and Stochastic Gradient MCMC give asymptotically correct samples of the posterior though scale to large models poorly. Some variational inference algorithms such as mean-field variational Bayes and variational dropout all give scalable yet biased approximations. Monte Carlo Dropout (Gal & Ghahramani, 2016)- a test time interpretation of dropout as a kind of approximate Bayesian inference, is a more convenient and popular approximation that has been used in medical imaging and other safety-critical systems.

More recently, deep ensembles (Lakshminarayanan, Pritzel, and Blundell, 2017) have proved to be a highly effective and computationally efficient way of quantifying uncertainties: when a bunch of networks are trained with random initializations, the ensemble can be used to estimate uncertainties in question very well, more than many more complex Bayesian approximations would over standard benchmarks. But ensembles have computational costs that are directly proportional to the size of the ensemble, which restricts their application in resource constrained deployment. Most recent efforts have attempted to realize the advantages of ensembles in a more efficient manner using hypermodel ensembles, batch ensembles, packed ensembles, and techniques to determine uncertainty using spectral normalization and feature-space density estimates (Van Amersfoort et al., 2020; Mukhoti et al., 2023). Recently, convergence has been given significant attention by conformal prediction through a variety of attempts to provide distribution-free prediction sets with valid marginal coverage guarantees, which is appealing as it places minimal conditions on the model structure or data distribution, and can be used with a given model after-hoc. Conformal prediction has been actively studied in recent years, merged with state-of-the-art deep learning systems and generalized to sequential, multi-step, and structured prediction tasks, under active investigation, and with important implication on safety-critical deployment.

### 3.4. The third kind will be Interpretability, Explainability, and Transparency

The fact that deep neural networks are opaque, and implement a black box that processes inputs to outputs by millions of non linear computations whose identifying intermediate representation cannot be expressed in human understandable form is a significant challenge to trustful deployment. Interpretability and explainability, though frequently used interchangeably, are two different phenomena: interpretability is the level at which the inner workings of a model can be comprehended by the human viewer, whereas explainability is the ability to give a post-hoc decision as to why one particular prediction of a model and not another can be made. The difference is important as it explains how it should be represented, the representations and calculations learned by neural networks can be reversed, producing interpretability research (mechanistic interpretability, circuit analysis, probing classifiers), whereas it can be described with faithful and human-understandable explanations of particular predictions, which is explainability research (LIME, SHAP, integrated gradients, attention visualization).

Interpretability has been brought to a knife by the appearance of very large foundation modelsGPT-4, Gemini, Claude, and their successors the billions of parameters, their emergent capacities, and complex reasoning behaviors are those that are especially hard to understand mechanistically. Mechanistic interpretability, an area of research, has

explored a program of finding the circuits present in transformer networks that support particular capabilities, with a significant amount of success in explanations of induction heads, indirect object identification and modular arithmetic, but a full mechanistic account of the capabilities of frontier models has yet to be achieved. More recent conceptual arguments have demanded that post-hoc methods like SHAP and LIME are faithful enough to model behavior to be trusted to various high-stakes settings: research has found that the methods themselves produce contrasting explanations across various measures of faithfulness, themselves are vulnerable to manipulation (Slack et al., 2020). Formulation of interpretation techniques fulfilling the requirements of simultaneously faithful, stable, human-understandable, and computationally manageable large-scale models is a key societal issue of trustworthy deep learning.

### 3.5 Equity, Discrimination and Society

The spread of social prejudices with deep learning systems is one of the issues that combine the technical and normative aspects of credibility. Deep learning models which are trained on historical data are bound to implement the structural inequalities within the historic data, and unless additional measures are taken, such encoded inequalities would result in discriminatory predictive models across demographic lines such as race, gender, age, disability, and socioeconomic status. The bias sources are many and interact complexly: sampling bias (training data that over- or under-represents groups), label bias (human created labels that are a biased reflection of annotator biases), representation bias (features that represent proxy variables of some attribute that is being protected), aggregation bias (models that average over heterogeneous subgroups to the detriment of minority groups). Various mathematical definitions of fairness, such as demographic parity, equalized odds, calibration within groups, individual fairness, and counterfactual fairness, different fairness intuitions, represent real philosophical disputes about distributive justice which cannot be settled using purely technical information.

The emergence of fairness-aware deep learning has investigated various interventions happening at training-time and post-processing layers, such as adversarial debiasing (training a predictor while adversarial fooling a discriminator that tries to predict protected attributes), fair representation learning, reweighting and resampling of training data, and constrained optimization formulations (that use fairness constraints). At the same time, the emergence of intersectional fairness models, which take into consideration the racial accumulation of disparities between two or more attributes, has shown that initiatives aimed at preventing or eliminating disparity on one line of protection, consequently, lead to extreme disparities at the intersections (such as a model able to create parity between men and women still with extreme disparities between old-aged Black women). The increasing literature at the intersection of algorithmic fairness

and causal inference provides promising theoretical basis of intervention methods to achieve fairness that is effective in respect to causal structure, but how to practically implement the frameworks of such approaches to large-scale deep learning systems is a major research problem.

## 4. Applications in the Fields that require Reliable Deep Learning

### 4.1 Healthcare and Clinical Decision Support

The field of healthcare is, possibly, the most challenging when it comes to relying on reliable deep learning, as the degree of decisions, regulatory aspects, distributional heterogeneity, and the necessity to avoid causing harm are closely interconnected. Healthcare Deep learning systems have a very long diversity of applications: pathology: slide analysis to identify and grade cancer, radiology: image interpretation to screen and stage the disease, clinical: natural language processing of electronic health records, drug discovery and molecular properties prediction, genomic: optimism interpretation, clinical: risk stratification and, more recently, multimodal large language models integrate image, text, and structured clinical data into reasoning systems.

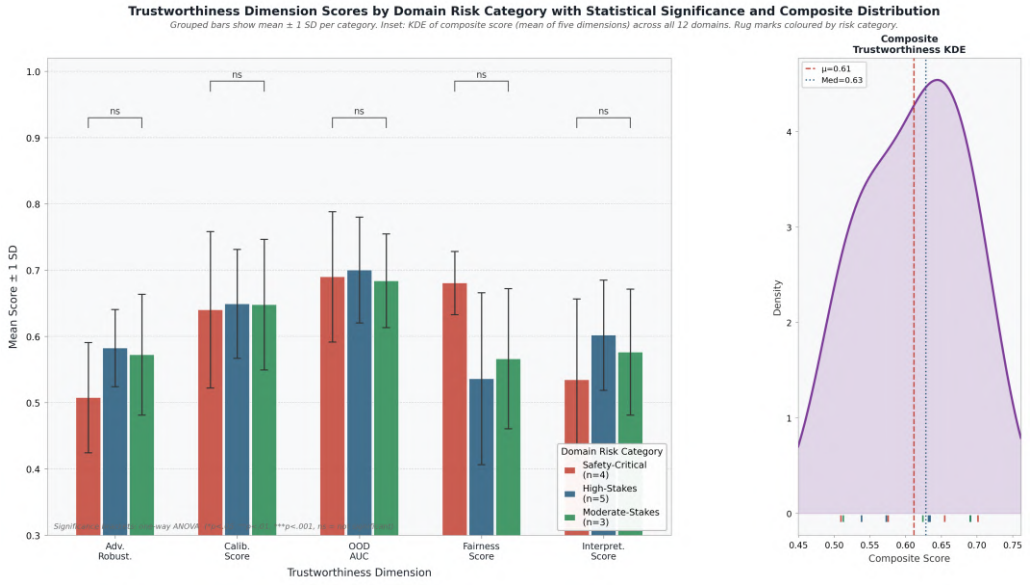


Fig. 3 Grouped Bar Chart with ANOVA Significance Brackets + Composite KDE Inset

Fig. 3 Compares mean ± 1 SD trustworthiness scores across Safety-Critical, High-Stakes, and Moderate-Stakes domain categories for all five dimensions. Significance brackets atop each group cluster report one-way ANOVA *p*-values using standard star

notation. The inset right-hand panel shows a KDE of the composite (mean-of-five-dimensions) trustworthiness score across all 12 domains, with rug marks colour-coded by risk category—providing a single-number summary distribution for rapid cross-domain comparison.

All of these applications face the entire range of trustworthiness challenges: training models using information of particular hospital systems, imaging equipment, and demographic characteristics may not work in other institutional settings; clinically pertinent uncertainty has to be presented in forms comprehensible to clinical users; regulatory compliance with FDA Software as a Medical Device guidance has to be demonstrated in written form demonstrating safety and effectiveness across diverse population groups; and deliberate propagation of health inequity by biased training data is an issue of concern in all applications.

The gradual development of the FDA with regard to AI/ML-based Software as a Medical device (SaMD) have sequentially narrowed its requirements of pre-market validation and post-market monitoring of deep learning diagnostic systems. Interestingly, the framework tackles the problem of the so-called adaptive or continuously learning AI systems that change their parameters once deployed a system, a situation that necessitates the presence of effective mechanisms to detect performance degradation, cause revalidation, and patient safety when changing models. The recent regulatory efforts have also been directed at the demographic performance reporting requirements: the FDA instructions regarding a more diverse clinical trial population on the ground have spillover effects to the demographics of the datasets to validate AI/ML-based medical devices, as developers should be able to describe and report performance statistics in appropriate subpopulations. Such regulatory changes are indicative of an increased recognition that benchmark performance on a held-out test set is a weak reflection of clinical utility and safety in the real world and that careful characterization of distribution shift, uncertainty and demographic performance is the key to responsible deployment.

### 4.2 Autonomous Systems and Robotics

The autonomous transportation systems, Unmanned Aerial Vehicles, and robotic manipulation systems can be viewed as the category of application where the consequences of failures in deep learning are instant, physical, and even fatal. The subsystems of autonomous vehicle perception, prediction, and planning, based on deep learning, include convolutional networks to detect camera-based objects, point cloud networks to understand the surrounding with LiDAR, multi-sensor fusion and motion prediction through transformer-based models, which confront all the issues of trustworthiness listed in this chapter. The safety case of autonomous vehicles should

further make clear that the cumulative failure frequency of the system in all operating conditions should be below a threshold comparable to an actual human driver, which makes comprehensive characterization of failure cases under distribution shift (uncommon weather conditions, sensor degradation, special traffic conditions), adversarial conditions (physical adversarial attack on vision, special response to adversarial conditions), and edge cases which may be grossly underrepresented in training data but are nonetheless decisive in practice, all a requirement.

Formal verification of neural networks has also advanced in terms of offering certified guarantees on the behaviour of neural networks trained on given input domains, and a variety of tools, including Reluplex, a-b-CROWN, and VeriNet, has been demonstrated to be able to verify properties of tens of thousands of neuron networks within realistic time constraints. Nonetheless, it is still not possible to scale certified verification to the large networks used in production autonomous systems, that is, with a large number of millions up to billions of parameters and on continuous and high-dimensional sensor spaces. The current practice in the industry has thus embraced defense in depth where runtime monitoring systems (to identify abnormal inputs and introduce predictions suitable to be reviewed by humans), the predictive ability of ensembles (Great care with prediction uncertainty) (to instigate conservative behavior when prediction uncertainty is greater than calibrated limits), comprehensive scenario-based testing procedures, and operational design domain constraints which limit deployment to operating conditions within the validated operating envelope. The enhancement of formal operational design domain speification, and the engineering equipment to confirm that the demanding behavior of a system stays within range in its specifications throughout the prescribed domain are an lively field of standards development in the automotive safety engineering (ISO 26262, ISO 21448 SOTIF).

### 4.3 Financial Services and Systemic risk.

The application of deep learning in financial services, such as credit underwriting, fraud detection, algorithmic trading, insurance pricing and systemic risk control, presents challenges of trustworthiness, determined by unique aspects of financial data and regulatory space. Financial data are highly concept drifting: the statistical relationships of features and outcome are varying with the changing market conditions, economic regimes, customer behavior, and fraud strategies. A credit score model that is modeled based on the data at a time of low interest rates and economic growth can fail in a catastrophic manner once the interest rates and default rates rise- failure mode directly reminiscent of the model risk events that helped triggered the 2007-2008 financial crisis. The systemic (in particular the complexity of institutions) aspect of financial AI can prove especially crucial in having correlated behaviors across many institutions stressing

the same models: this can reinforce instead of dissipate shocks, making the system vulnerable to systemic fragility that cannot be identified or mitigated at the level of individual model risk management.

Financial deep learning has regulatory needs that are much higher than those of other low-regulatory paradigms, affecting its interpretability. The United States Fair Credit Reporting Act and the Equal Credit Opportunity Act demand that negative credit actions be clarified to a candidate in terms of particular measles-backing reasons- a need that the cloudy deep learning models cannot meet and has motivated the creation of locally faithful clarifications as well as the establishment of them in model management systems. This is shown by the SR 11-7 recommendations on the Federal Reserve Board on the Model Risk Management that sets the expectations of model validation, stress testing, and continuous monitoring, and extends them to machine learning models, as it requires presentation of evidence concerning model reliability in adverse conditions. The Basel III and developing Basel IV regimes have capital requirements which are hedged against model uncertainty, and which provide explicit financial reward to the better quantification of uncertainty in risk models. The point of convergence between adversarial robustness and financial stability is also becoming acute: adversarial attacks on fraud detection systems, in the form of creating transactions that are not detected, and model extraction attacks, in the form of grabbing proprietary model parameters, are each a serious operational risk of financial institutions.

### 4.4 Large Language Models and Natural Language Processing.

The recent advent of large language models (LLMs) as general-purpose AI systems, which can understand natural language, generate it, reason, produce code, and multimodal process languages, has created a new qualitatively new range of trustworthiness issues that both scale and restructure the issues that led to the study of trustworthy deep learning in more limited contexts. The production of factually erroneous, internally inconsistent, created example text fluently and with confidence (also known as hallucination), is a failure mode that has nobler failures in classification and regression systems but is particularly debilitating since it is hard to recognize without prior domain knowledge. LLMs are seriously problematic in their calibration: models are not inherently giving calibrated probability probability over free-form text output, and the relation between verbal statements of confidence (I am confident that... ) and accuracy is weak and untrustworthy. The issue of sycophancy, which is the tendency of LLMs to amend their proclaimed beliefs when pressurized by users instead of evidence, casts the question of whether the results of the thinking performed with the help of LLMs are reliable in epistemological terms.

Adversarial robustness as adversarial injection-based challenges to the trustworthiness of LLMs such as adversary injection attacks, where spearheaded by malicious content in the context of a model, the model is compromised into acting in ways contrary to expectations; jailbreak attacks, where a carefully crafted prompt and then placed in an agentic setting within a natural language text or web page administrates a model to behave in manners unhelpful to the processing; and indirect prompt injection, where a malicious command is embedded in texts, web pages, or tool output which the model is used in, and Safety and alignment of LLMs is viewed as a separate research field of trustworthy AI including reinforcement learning from human feedback (RLHF), constitutional AI, direct preference optimization (DPO), and more recent approaches to scalable oversight and debate, which aim to preserve alignment as models continue to scale up. The application of LLMs to agentic environments, where the systems act in sequences of actions that have real-life implications by using tools, executing code, or making API calls, dramatically increases the stakes of failures to be trustworthy and encourages new study on the topic of uncertainty-aware planning, reversibility constraints, and human oversight procedures to agentic AI systems.

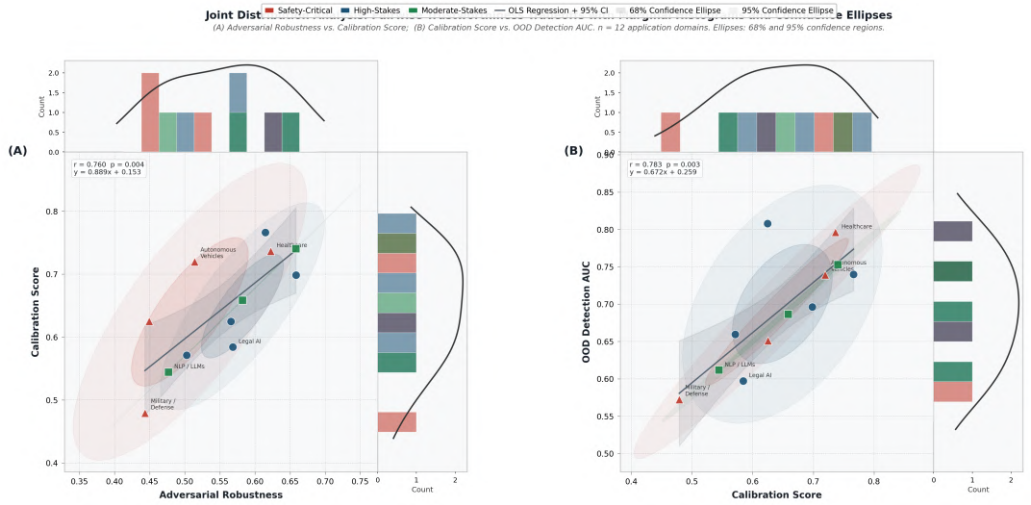


Fig. 4 Joint Distribution with Marginal Histograms and Confidence Ellipses

In above Fig. 4 Two side-by-side joint distribution panels: (A) Adversarial Robustness vs. Calibration Score, and (B) Calibration Score vs. OOD Detection AUC. Each panel features domain-annotated scatter points, 68% and 95% covariance confidence ellipses per risk category, a global OLS regression line with 95% CI, and Pearson *r* with regression equation. Marginal panels show stacked histograms with KDE overlays for both axes, revealing the marginal distribution shape of each pairwise combination—

directly illustrating the robustness–calibration and calibration–OOD tradeoff structures discussed in the chapter.

### 4.5 Science Discovery and Essential Infrastructure.

Deep learning is being applied in more scientific discovery situations such as in materials science, climate modeling, drug discovery, genomics, and particle physics: model predictions must have a high level of trust in order to be scientifically valid and ultimately whether scientific conclusions about policy and medical practice may have a real-world impact. The prediction of protein three-dimensional structure based on an amino acid sequence by AlphaFold2 (jumper et al., 2021) is perhaps the most popular current achievement of deep learning in scientific discovery and also demonstrates the challenges in trustworthiness in the field: the model works correctly on proteins that are similar to the training set but makes major errors on proteins in new protein families, membrane-bounding proteins, and disordered regions that are hard to predict originating thereof--a trend in performance that needs effective uncertainty measurements to be reported to users. Physical consistency of climate modeling with deep learning Cross-laminated neural network surrogates of costly numerical simulation aspects must be not only consistent with the data distribution of the training cases but must obey physical conservation principles, numerical stability, and interregional extrapolation properties at, or exceeding, climate regimes that were not observed during the training period.

Critical infrastructure systems such as power grids, water treatment systems, transportation systems, and telecommunications are a field where the intersection of the deep learning with operational technology presents security and safety hazards of utmost importance. The use of deep learning to detect anomalies in industrial control systems (ICS), to predictive maintain, and to optimize controls creates additional attack surfaces to exploit, an attacker who can engineer inputs that trigger an anomaly detection system to overlook an actual attack signature or an attacker who can engineer inputs that cause a control optimization system to empower destabilizing actions can have results akin to a physical attack on the infrastructure. Nation-state adversaries with advanced capabilities have been reported to target ICS with growing aggressiveness and there has been a systemic risk of unprecedented kind of attack against vulnerable AI components into critical infrastructure system without proper adversarial resiliency testing, which dominates scholarly discussions.

Table 1: Motivations, Challenges, and Risk Profiles Across Key Application Domains of Trustworthy Deep Learning

| Application Domain | Trust Motivation | Primary DL Challenge | Risk Category | Illustrative Example |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Healthcare & Clinical AI | Patient safety, liability, regulatory compliance | Distribution shift, spurious correlations | High-stakes, life-critical | Misclassification of malignant tumors due to dataset bias |
| Autonomous Vehicles | Legal accountability, real-world unpredictability | Sensor adversarial attacks, OOD detection | Safety-critical, physical harm | LiDAR spoofing causing collision in self-driving car |
| Financial Systems | Market stability, regulatory mandates (Basel IV) | Concept drift, model opacity | Economic, systemic risk | Credit scoring model amplifying demographic disparities |
| Cybersecurity & Intrusion Detection | National security, data integrity | Adversarial evasion, poisoning attacks | National security, privacy | Malware disguised as benign code to evade DL detector |
| Natural Language Processing | Fairness, misinformation prevention | Hallucination, prompt injection | Societal, epistemic harm | LLM generating confidently wrong medical advice |
| Facial Recognition & Biometrics | Civil rights, algorithmic fairness | Demographic bias, adversarial patches | Legal, discrimination risk | Misidentification of dark-skinned individuals by surveillance AI |
| Drug Discovery & Genomics | Scientific reproducibility, patient outcomes | Small data regimes, uncertainty calibration | Research integrity, patient safety | Molecular property prediction failure due to covariate shift |
| Critical Infrastructure | Operational resilience, national security | Byzantine fault tolerance, targeted attacks | Physical and economic harm | Power grid control system manipulated via adversarial inputs |
| Legal & Judicial AI | Due process, equal protection | Unexplainability, demographic bias | Constitutional, human rights | Recidivism prediction tool exhibiting racial bias |
| Educational Technology | Equity, developmental impact on minors | Feedback loop reinforcement, overfitting | Developmental, pedagogical harm | Adaptive tutoring system entrenching learning gaps |
| Climate & Environmental Modeling | Policy accuracy, public trust in science | Epistemic uncertainty, model misspecification | Global governance risk | Sea-level prediction errors distorting adaptation policy |
| Military & Defense AI | Laws of armed conflict, proportionality | Adversarial robustness, sensor deception | Existential, ethical catastrophe | Target classification error in autonomous weapon systems |

## 5. Exploring the Technical Landscape: Approaches and Methods

### 5.1 An Integrated Approach to Reliable Quality

The abovementioned multiplicity of technical challenges could indicate that reliable deep learning is a collection of problems that are not clearly articulated and do not have distinct toolkits. In practice, though, the dimensions of trustworthiness interrelate with each other profusely, and a single framework that places them as complement to each other in what is fundamentally one underlying concern, which is the validity of model performance to the values and expectations of human beings in the full variety of deployment conditions is both conceptually insightful and practically desirable. In this respect, the common issue in all problems of trustworthiness is the problem of generalization in uncertainty: the problem of guarantee behavior by a model, based on finite, biased, possibly adversarially corrupted data, in the open-ended, changing, adversarial conditions of actual implementation.

Such a coherent thinking implies a research agenda that would advance on the three complementary fronts. To begin with, how training algorithms and model structures can create trustful characteristics into the training model by default, i.e., adversarial training, training to invariant properties, uncertainty-sensitive architecture, equity objectives, and interpretability regularization. Second, the creation of methods from post-hoc that can be used to enrich trained models with trustworthy properties, such as calibration, conformal prediction, generation of explanations, and formal verification. Third, enhancement of deployment infrastructure and governance that uphold trustful conduct in operational settings characterized by dynamism such as runtime monitoring, distribution shift identification, continuous evaluation frameworks, human-AI interaction frameworks to adequately distribute decision-making authority among mechanized systems and human control. These three fronts actively interact with each other, and the chapters that follow this book elaborate on each of them.

### 5.2 Novel Research Frontiers and Trends

The technical landscape of the trustworthy deep learning is being re-formed or reconfigured by several emerging trends that create a new challenge and bring a new opportunity. The Abstract of foundation models and transfer learning has created a paradigm shift in the model development process: instead of more direct training of models to particular tasks, more practitioners are increasingly fine-tuning massive pretrained models on task-specific data. The change in paradigm has far-reaching consequences regarding the concept of trustworthiness: the distributional biases, adversarial vulnerabilities, and safety failures that are transferred by fine-tuning to

downstream tasks can be encoded by the foundation model, and the assessment of trustworthiness should now consider the base model, the training of the fine-tuning task, and the final task-specific model as the whole system. Simultaneously, foundation models offer novel opportunities regarding the representational richness, which then leads to zero-shot and few-shot adaptation to new domains, which might cause fewer distributional shifts of downstream applications.

Neuro-symbolic integration integration, which is developing the statistical learning properties of deep neural networks with the logical reasoning and knowledge modeling properties of symbolic AI systems, is one approach to making AI systems stronger, more interpretable, and also more calibrated in fields where knowledge is a constrained resource. Neural perception systems coupled with symbolic reasoning and probabilistic inference can, in principle, be able to explain things by a set of logical rules and domain knowledge, offer calibrated uncertainty by applying probabilistic inference, and to do this compositional generalization to novel combinations of known concepts-solving a number of the trustworthiness problems of purely subsymbolic systems. This potential is actively being implemented in scalable neurosymbolic architectures that are as fast as large-scale purely neural methods and which pose a fast-paced and highly dynamic future research frontier. At the same time, the development of formal methods and verification techniques of neural networks, including SMT-based verification, abstract interpretation, and neural network certification toolboxes, gives an opportunity to have a mathematical safety guarantee of neural network components in safety-critical systems, at computational costs that remain untenable at present to small-scale networks or to highly specialized specifications.

What interests in particular attention is the convergence between privacy-conserving machine learning and trusted deep learning as one of the many frontiers. A mathematical model named as differential privacy (DP), which ensures that the result of an algorithm does not give significantly more information about a particular individual than the outcomes would have given were the data of that very individual not available, offers a principled method of training deep learning models respectful of the privacy of data subjects. Federated learning (FL) allows the training of models on distributed datasets with no data centralization deployed to expose sensitive information, and gaps adherence to data residency constraints. Yet, privacy, robustness, fairness, and accuracy are not exactly compatible, as: introducing noise in the form of differential privacy in training may decrease the model accuracy and may magnify the difference in demographic performance between participants; federated learning systems may be susceptible to a malicious model poisoning attack; and the privacy-utility tradeoff is an essential limitation that has to be negotiated very carefully in practice. Proposing training algorithms that both reach high privacy, high robustness, fairness, and accuracy are achieved is also a large research problem with important practical impacts.

Table 2: Key Techniques for Trustworthy Deep Learning — Dimensions, Theoretical Bases, and Deployment Status

| **Technique / Method** | **Trust Dimension Addressed** | **Theoretical Basis** | **Computational Cost** | **Maturity & Deployment Status** |
|---|---|---|---|---|
| Adversarial Training (PGD, TRADES) | Adversarial Robustness | Min-max optimization on worst-case perturbations | Very High (3–10× standard training) | Mature; widely deployed in security-critical models |
| Certified Defenses (Randomized Smoothing) | Adversarial Robustness | Probabilistic certification via Gaussian noise | High (requires Monte Carlo sampling) | Research-stage; adopted in safety-critical benchmarks |
| Deep Ensembles | Uncertainty Quantification | Variance across independently trained networks | High (scales linearly with ensemble size) | Widely deployed; de facto standard for UQ |
| Monte Carlo Dropout (MC-Dropout) | Uncertainty Quantification | Variational inference approximation via dropout | Low-Moderate (inference-time sampling) | Very mature; production use in medical imaging |
| Bayesian Neural Networks (BNNs) | Uncertainty Quantification | Posterior inference over network weights | Very High (MCMC or variational methods) | Research; scalability remains an open challenge |
| Conformal Prediction | Uncertainty Quantification | Distribution-free coverage guarantees | Low (post-hoc calibration) | Rapidly growing; recent regulatory interest |
| Data Augmentation & Invariant Risk Minimization | Distributional Robustness | Causal representation / domain generalization | Moderate | Mature for CV; emerging in NLP and tabular data |
| Certified Robust Training (IBP, CROWN) | Certified Robustness | Interval bound propagation / linear relaxations | Very High | Research; emerging industrial use in verification tools |
| Gradient Masking & Input Purification | Adversarial Resilience | Preprocessing-based defense layer | Low-Moderate | Partially deprecated; used as auxiliary defense |
| Explainability-Constrained Training (CDEP, RRR) | Robustness + Interpretability | Penalizing reliance on spurious features | Moderate | Emerging; active research area since 2022 |
| Federated Learning with Differential Privacy | Privacy + Distributional Robustness | Local differential privacy + aggregation protocols | High (communication overhead) | Deployed in mobile (Gboard, Apple), healthcare |
| Test-Time Adaptation (TENT, TTT++) | Distributional Robustness | Entropy minimization / self-supervised adaptation | Low-Moderate (inference-time update) | Emerging; strong results on corruption benchmarks |

## 6. Trustworthy Deep Learning Sociotechnical Dimensions of Trust

The definition of trustworthiness and the possibility of its realization is placed into the context of trusted deep learning, which has the technical research agenda embedded in it. Trustworthiness does not constitute a technical feature of a model or an algorithm, but it is a relational, that characterizes the aptness of the trust that humans and institutions put on an AI system within a certain context of application. Even a model that satisfies the technical criteria of robustness, calibration and fairness might be unreliable in the context where its results are going to be abusively interpreted, or where the humans who are going to use it are not expert enough to assess its advice critically, or where the institutional motivations according to which the model is used are inconsistent with the interests of the populations that it influences. On the other hand, a model that is technically constrained can be deployed in a responsible and reliable manner in a well-planned sociotechnical system that deals with proper human supervision and interpretive scaffolds and feedback systems.

The existence of trustworthy AI is progressively known to have a significant role on its human aspect as a determinant of real-life outcomes. Studies of algorithm aversion and algorithm appreciation have recorded the many and context-sensitive manner of how human decision-makers react to AI recommendations: some respond with over-reliance on AI recommendations (automation bias), wherein the model is more accurately declared to be wrong); other respond with systematic discount of AI recommendation (algorithm aversion), wherein the model is seen to offer benefits of genuine accuracy that it can be delivering to human decision-makes. Human-AI interaction system designs that encourage proper reliance-violent to real model performance and uncertainty is currently being actively investigated at the cutoff point of cognitive science, human-computer interaction and trustful AI. Meaningful human control formulated in the ethics and governance of autonomous systems offers a normative framework on specifying what should be sufficient human control of AI decisions based on the stakes of the decision, human expertise, and system reliability.

The institutional and organizational aspects of trustworthy deep learning involve creating model governance procedures, audit procedures, documentation requirements, and accountability mechanisms of the embedding of technical trustworthiness requirements within organizational procedures. There are early attempts to make model capabilities, constraints, and conditions of intended use in documentation standardized in model cards (Mitchell et al., 2019) and datasets datasheets (Gebru et al., 2021) that need to be conveyed to enable knowledgeable downstream decision-making by deployers, regulators, and populations impacted. The new form of algorithmic auditing systematic analysis of the work of deployed AI systems by third parties regarding their performance, fairness, and safety is the reflection of the already existing practice of financial one that establishes an external sense of responsibility in the organizations that

mature consequential AI systems. Policy to fully develop an algorithmic auditing audit standards, methodology, and professional infrastructure is an active field of policy development that has important ramifications concerning the practical implementation of trustful algorithmic scale of auditing.

## 7. Conclusion

This chapter has explored that reliable deep learning is not an incidental or optional issue but is a prerequisite that the conscientious educational use of deep learning is put into practice in the entire gamut of areas where the capabilities of the latter can produce actual social utility. The incentive reasons behind trustworthiness can be summarized as multiple and mutually reinforcing: the empirical experience of failures in the real world, the growing pressures of regulation, the epistemological inadequacy of conventional empirical risk minimization, and the ethical demands of justice and accountability all lead to the conclusion of the inadequacy of the accuracy measures that dominate academic benchmarking as proxies of the reliable, safe and fair applications that are required of high-stakes applications. These technical difficulties that should be surmounted are daunting, they are adversarial vulnerability, distributional shift, uncertainty miscalibration, opacities, and bias all represent profound structural constraints of the existing deep learning systems, which are difficult to combats and interact to form complex systems. The fields of application these issues are most acute, including healthcare, autonomous systems, finance, natural language processing, scientific discovery, as well as critical infrastructure, are also exactly the areas where the benefits of deep learning have the most potential and where their drawbacks are the most recreational.

Trustworthy deep learning is not only about using some available technical tools as solutions to new challenges; it is about creating entirely new theoretical and algorithmic frameworks and appraisal practices that do not made reliability, safety, fairness and accountability but actively design these aspects in practice. It also needs to look into the sociotechnical aspects of AI implementation- the human aspects, institutional structures as well as governance mechanisms that determine how the technical features can be transformed into practical consequences. The merging of technical precision, moral obligation and regulatory observance should not be an area of tension but what it comes down to: the challenge of creating AI systems that are both competent and trustworthy is the challenge of scientific and engineering achievement of our time, and achieving it will need the convergence of both fields. This book is made as part of that endeavor.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 2: Fundamentals of Deep Neural Networks and Reliability Issues

## Abstract

Deep neural networks (DNNs) have fundamentally changed the viewpoint of research and practice in the field of artificial intelligence to create breakthroughs in the field of computer vision, natural language processing, autonomous systems, scientific discovery and so on. Nevertheless, in addition to superb empirical achievements, deep learning systems have a segment of severe reliability deficiencies that describe their safe and reliable use in high-stakes applications. The chapter gives an extensive background approach to the deep neural network structures and the multidimensional reliability concerns they pose. Starting with the mathematical and computational foundations of the modern neural architecture, the chapter discusses the main ways in which neural architectures fail, such as adversarial vulnerability, distributional brittleness, calibration issues, uncertainty quantification issues, privacy leakage, backdoor vulnerability, shortcut learning, and catastrophic forgetting, in a systematic manner. New architectures and the most recent advances of the trustworthy AI between 2023 and 2025 are incorporated across the board. The synthesizing summary tables are also provided to make it easier to conduct a comparative evaluation of architectures and reliability dimensions. The chapter sets the conceptual terminology and technical base on which later chapters develop and by doing so puts reliability no longer as a design consideration but a first-order design requirement in the system of deep learning.

## 1.Introduction

This is a key process that has led to the emergence of an immediate necessity, namely to ensure that not only the neural networks themselves but also their application to practice, in all its variations and varieties, are considered reliable, consistent and understandable when applied to the entire range of real-life situations. A very remarkable scaling of model capacity has been seen in the field, with the number of parameters increasing over

time, as has been seen in AlexNet (2012) up to hundreds of billions of parameters, in modern large language models, and in multimodal foundation models. Nonetheless, this scaling has not produced scale-proportional improvements in robustness, calibration, and sensitivity to distribution shift, instead numerous reliability issues have compounded with complexity of the model (Bommasani et al., 2021; Hendrycks et al., 2023). This paradox unfolding capacity and constant vulnerability spurs a whole sub-discipline of reliable deep learning which in turn has quickly become the focal point of the entire academic research community and regulators across the globe.

The deep learning meaning of reliability is multidimensional in nature. A model can be highly average in all aspects whilst disastrously failing on underrepresented subpopulations, attentively near decision boundaries, spillage sensitive training data through model query interfaces, or make convincingly sounding, yet actually false outputs. All these failure modes are indicative of unique structural weakness of modern DNN architectures and each of them requires a corresponding set of diagnostic tools, mitigation strategies, and evaluation procedures. The study of these failure modes needs, on the one hand, a clear understanding of the construction of deep neural networks, training, and generalization, which even with all its apparent familiarity bring completely fundamental surprises at the edge of fundamental investigation of theory and practice.

This chapter has the following organization. Section 2 gives the mathematical background of both the feedforward and deep architectures, including the forward and backward propagation algorithms, choice of activation functions, and optimization landscapes. Section 3 gives a survey of the major architecture families currently in use, such as convolutional, recurrent, attention-based and generative architecture, considering their structural inductive biases and failure modes. Section 4 offers a methodical discussion of the most significant areas of reliability failure in deep learning systems that are arranged by the cause and action of any failure mode. Section 5 discusses cross-cutting themes in trustworthy AI, such as their new regulatory and principled design paradigms of robustness-by-construction. In section 6, the two extensive summary tables are provided. Section 7 of the chapter is a synthesis of the chapter, which looks ahead.

## 2. Deep neural network mathematical foundations.

### 2.1. The Feedforward Computer Graph.

A deep neural network can be formally defined as a parametric function f_θ: ℝ^d → ℝ^k, composed of L sequential layers, where each layer l applies an affine

transformation followed by a pointwise nonlinearity: $h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)$. The parameters$\theta = W^{(l)}, b^{(l)}{}_{l=1}^{L}$ are learned by minimizing an empirical risk functional over a training dataset D = {(x_i, y_i)}_{i=1}^{n}, typically of the form L(θ) = (1/n) Σ ℓ(f_θ(x_i), y_i), where ℓ is a task-appropriate loss function such as cross-entropy for classification or mean squared error for regression. The theorems of universal approximation of Cybenko (1989) and Hornik (1991), and universal approximation extended to deep architecture by Lu et al. (2017) among others, allow one to say that wide single-hidden-layer networks can approximate any continuous function on not too small supports, but they do not necessarily say whether such approximations would be statistically or computationally efficient.

The nonlinear activation, s, takes the middle stage in determining the expressivity of the neural networks as well as their stability. The rectified linear unit (ReLU), which is s(z) = max(0, z), has become the default in place after popularization by Glorot et al (2011) because it alleviates the vanishing gradient problem and the sparsity of its activations. Nevertheless, ReLU also has its own pathologies: it is non-differentiable at zero, can experience the so-called dying ReLU effect where neurons always output zero, and activations in the positive half-plane which are linear, subjecting the representational power of shallow architectures. Some variants like Leaky ReLU, Parametric ReLU, Exponential Linear Units (ELU), Swish (s(z) = z*sigmoid(bz)) and the Gaussian Error Linear Unit (GELU) have been suggested to overcome these drawbacks. GELU, a modern transformer component applied to BERT and GPT-4, can be seen as an approximation of a stochastic regularization mechanism, in which inputs are weighted by their quantile under a standard normal distribution, and has been demonstrated to generate better-calibrated representations on a number of NLP benchmarks (Hendrycks and Gimpel, 2016). The latest activation paradigm to be presented in the literature is Kolmogorov-Arnold Network (KAN) activation, where the stationary activation functions are parameterized by learnable B-spline functions on the network edges instead of nodes, projected on the network edge, and this has better interpretability as well as an enhanced parameter efficiency in certain settings of function approximations.

### 2.2 Landscapes and Generalization Optimization.

A deep neural network can be trained by moving around a non-convex loss-scape in high dimensions. The most common processes by which the parameters th are updated include stochastic gradient descent (SGD) and its adaptive forerunners, namely, Adam (Kingma and Ba, 2015), AdamW, Adadelta, and new optimizers AdaFactor and Sophia. One of the core theoretical issues in the field of contemporary deep learning is the seemingly counterintuitive phenomenon of the non-convexity of the loss landscape and the overwhelming empirical success of deep learning training. According to the works

by Choromanska et al. (2015), Dauphin et al. (2014), and Du et al. (2018), registering the saddle points and local minima at high losses is uncommon in overparameterized regimes, and instead gradient descent discoveries solutions in wide and flat basins of the loss landscape. Such geometry has been found to be days with regard to reliability: convergent-to-sharp-minima models are poorly generalizing and have increased sensitivity to adversarial perturbations, whereas flat minima solutions have been known to be robust and calibrating (Foret et al., 2021). Sharpness-Aware Minimization (SAM) is an operationalization of this observation where loss sharpness is reduced together with loss, and, with standardized improvements in generalization and adversarial robustness across benchmark tasks, was proposed by Foret et al. (2021) and further expanded by much of the subsequent literature.

The deep neural network generalization behavior is (so far) only partially understood, especially in the overparameterized regime, where the number of parameters is much larger than the one of training examples. Classical statistical learning theory, based on VC dimension and Rademacher complexity predicts poor generalization at this regime but practice at this learning rate defies the prediction, which is called the problem of benign overfitting (Bartlett et al., 2020). More up-to-date theories, such as the neural tangent theory, mean field theory, and PAC-Bayes bounds, can all be used to explain why over-parameterized networks generalise, but none can thus far give a complete or predictive explanation. Empirically reported by Belkin et al. (2019) and theoretically researched since, double descent curves exhibit that the test error can decrease after crossing the classical interpolation threshold, allowing the standard bias-variance tradeoff account untangled. These theoretical gaps do not simply exist on paper: such a lack of ability to make generalization behavior theoretically predictable renders it essentially impossible to determine the reliability of a deployed model, which inspires empirical robustness benchmarking and uncertainty quantification protocols, which shall be discussed in later sections.

## 3. Modern Deep Learning Architectural Landscape.

### 3.1 Convolutional and Residual Architectures

Convolutional neural networks (CNNs) take advantage of the spacing arrangement of image-style data by common-weight convolutional filters that calculate local receptive field reactions, succeeded by hierarchical pooling functions that accumulate spatial data by accumulating semantic abstraction. Residual connections in ResNet as introduced by He et al. (2015) was an innovation that made it possible to train hundreds of layers in

networks without performance degradation, as it offered gradient highways where nonlinear transformations may be avoided, and gradient vanishing effects are reduced.

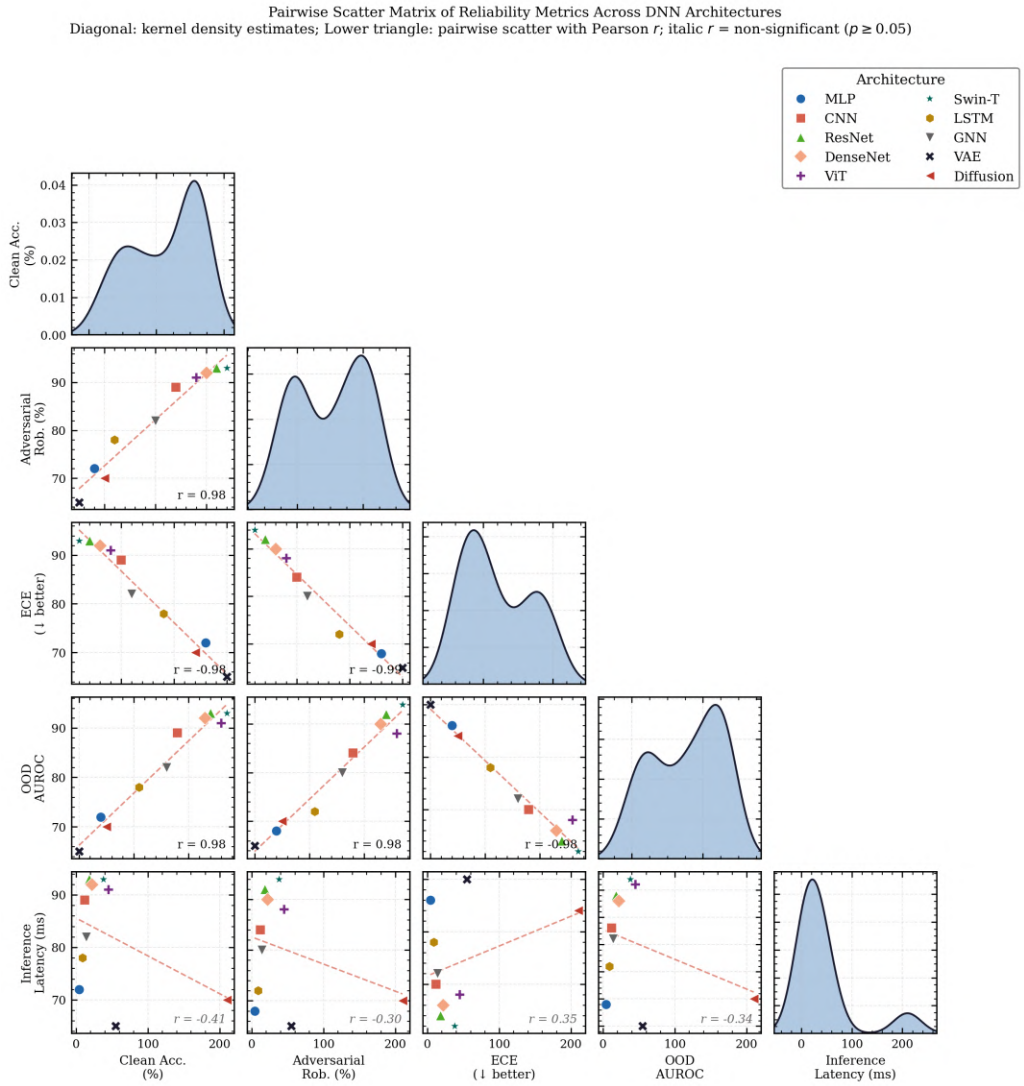


Fig. 1 Pairwise Scatter Matrix — Architecture vs. Reliability Metrics

Fig. 1 Shows inter-metric correlations across 10 DNN architectures on five reliability dimensions: Clean Accuracy, Adversarial Robustness,, ECE (calibration error), OOD-AUROC, and Inference Latency.

This architectural design has since been generalized into dense connections (DenseNet), inverted residuals (MobileNet) and compound scaling (EfficientNet). ConvNeXt architecture, introduced by Liu et al. (2022) and extended to ConvNeXt V2 (Woo et al.,

2023), proves that convolutional architectures still can (and in certain cases, outperform) vision transformers, suggesting that the common belief in the superiority of attention-based architectures to visual representation learning is invalid.

In spite of their excellent empirical performance, the CNNs have a well-known texture bias: unlike human vision, they tend to use local texture statistics, instead of global shape, to classify images, which is susceptible to naturalistic changes of distribution, whereby textures change and shapes do not. This texture favoritism is not a surface anomaly, but one which has one of the rich structural properties of convolutional feature hierarchies, and some partial solutions such as style transfer augmentation, shape-biased training, and vision transformers capturing long-range dependencies have been experimented with mixed success. Also connected with texture bias is the concept of shortcut learning- a property of CNNs that makes them utilize spurious relationships in their training sets in place of learning the causal or semantically engaging features that they will generalize on. Visible cases are models that use metadata of hospital scanners to categorize chest X-rays, as opposed to pathological characteristics, or image categorisers that use the background instead of foreground objects. Such failures are insidious in that they are not easily detected based on standard measures of accuracy, and can only be noticed as a result of a shift in distribution or a purposely constructed diagnostic test.

### 3.2 Transformer Architectures and Attention-Based Architectures.

Since introduction by Vaswani et al. (2017) in machine translation, the transformer architecture has since permeated the dominance of practically all modalities of machine learning, such as language (GPT-4, LLaMA-3, Gemini), vision (ViT, DINOv2, SAM), multimodal reasoning (GPT-4V, LLaVA, Flamingo), and code generation (GitHub Copilot, Code Llama). The scaled dot-product attention mechanism, Q KT / [?]) allows the attention of each token to be effective on all the other tokens in the sequence, modeling long-range dependencies that are difficult to efficiently model using convolutional and recurrent architecture. Multi-head attention is a generalization of this, calculating several attention patterns in parallel with learned subspaces, enabling the model to attend to various features of the input context at the same time.

Huge language models trained using transformer architecture have a complex of reliability issues that are qualitatively different to discriminative models. Such source of reliability crisis Hallucination also referred to as factually incorrect, internally inconsistent, or completely generated information with a high level of confidence has become the hallmark reliability crisis of modern large language models, and can be linked to the autoregressive training task that promotes fluent next-token prediction instead of grounding in facts or logical consistency. Sycophancy, or the effect of

instruction-tuned models to agree with the beliefs of users and modify outputs in accordance with the perceived user preferences as opposed to truth-telling, is a minor yet major alignment failure connected to the dynamics of reinforcement learning using human feedback (RLHF) optimization. Instantaneous injection attacks, where system instructions are bypassed or confidential information disclosed by adversarially constructed inputs, are a security vulnerability that is unique to the natural language interface paradigm. Large language models have a terrible characterization of calibration and tend to be highly task-dependent, and the models often show certainty in relation to frequency of training data, instead of being epistemically justified (Kadavath et al., 2022; OpenAI, 2023).

### 3.3 Generative Architectures and their Reliability Problems.

Generative models, such as variational autoencoders, generative adversarial networks, and diffusion models, have been capable of producing impressive realistic images, audio, video, and text and have also created new and more dangerous reliability issues. Tasers like generative adversarial networks have been proposed by Goodfellow et al. (2014), the network uses a minimax game between a generator G and a discriminator D, and is trained to reach a Nash equilibrium, where G generates the same samples as the data distribution. The GAN training process is in practice also infamously unstable, sensitive to mode collapse in which the generator is just trained to sample a limited set of modes of the target distribution, and prone to memorising training examples, which is a privacy risk of extreme importance when training data contains sensitive personal information. Several variants of the StyleGAN family (Karras et al., 2020, 2021) and the projected GAN variants have significantly enhanced the quality and diversity of the samples, yet the basic training stability and mode coverage issues remain.

Diffusion models which are trained to invert a denoising process applied stepwise over training examples have replaced GANs as the new paradigm of high-fidelity image synthesis after the work of Ho, Song, and Rombach et al. (2020) introduced Latent Diffusion Models, which consume much less computation since they are trained in a compressed latent space (and are therefore called implicit models). Diffusion models have a significantly better mode coverage and training stability than GANs, although they present entirely different reliability issues: their iterative denoising process is slow, which presents constraints to real-time applications, they suffer adversarial conditioning where engineered text prompts will cause harmful or copyrighted images; and they underlie the emerging problem of AI-generated content proliferation and associated challenges of provenance tracking, watermarking, and deepfake detection. The latest phenomenon that is of serious interest is the so-called model collapse, reported by Shumailov et al (2024) where generative models that are being trained sequentially on

their own synthetical output develop increasing levels of decline in quality and variety and threaten the validity of succeeding training pipelines that inadvertently include AI-synthetized content.

## 4.Deep Neural Networks Systematic Reliability Problems

### 4.1 Adversarial Vulnerability

Adversarial examples Adversarial examples are simply natural examples that have had purposefully introduced perturbations to achieve misclassification when fed to a deep learning model, the first to be systematically described by Szegedy et al. (2013), and have since been one of the most actively investigated phenomena in the field of deep learning. The foundational insight is that the high-dimensional input space of deep networks contains structured directions along which the loss function changes sharply, enabling an adversary with knowledge of model gradients to construct perturbations $\delta$ satisfying $\|\delta\|_p \leq \varepsilon$ (for small $\varepsilon$ under various Lp norms) that reliably cause targeted or untargeted misclassification. The Fast Gradient Sign Method (FGSM) of Goodfellow et al. (2015) and the iterative generalisation Projected Gradient Descent (PGD, Madry et al., 2018) are canonical attack benchmarks, and AutoAttack (Croce and Hein, 2020) is a useful parameter-free evaluation protocol that consists of a combination of complementary attacks, to prevent the overoptimistic robustness estimates that are remembered with competing attacks. The most empirically validated defense is the adversarial training scheme of Madry et al. (2018), which has adversarially perturbed examples as part of the training process through a min-max objective, and its cost in relation to the computer cost is not as high as later methods, or it causes a decrease in clean accuracy.

There is a parallel series of physical world attacks, using perturbations to printed object, road signs or clothing that can survive being photographed and reliably be deceived by deployed vision systems, creating immediate safety issues to autonomous automobiles and surveillance systems. In more recent literature, adversarial attacks have been considering the natural language domain, where character-level replacement, paraphrase based attacks, and universal adversarial triggers can instigate catastrophic failures in large language so far as sentiment classifiers and machine translation systems. Randomized smoothing (Cohen et al., 2019), Lipschitz-constrained architectures, and interval bound propagation have both developed certified defenses which offer provable guarantees that no perturbation within a given threat model can provoke misclassification, but attain certified accuracy at the cost of a fundamental tradeoff

between robustness and normal performance that only partially explains the state of this art.

### 4.2 The fourth section is titled Distribution Shift and Out-of-Distribution Robustness

In the real world deployment scenarios, there will always be evidences of statistical distributions changes in training and test data. Such changes are caused by natural differences in lighting, the behavior of sensors, geographic situations, time-dependent changes, or by imposing the systematic underrepresentation of rare yet consequential subpopulations in training data. A standard measure of natural robustness has become ImageNet-C benchmark (Hendrycks and Dietterich, 2019), a measure of model robustness against fifteen categories of algorithmically applied corruptions, such as noise, blur, weather, and digital distortions. Models which work well on clean ImageNet frequently catastrophically fail on these corruptions, and clean accuracy does not yield much information about corruption robustness, which is one reason why corruption robustness should be evaluated as a first-class model selection criterion. This paradigm is further extended to a wide range of real-world distribution shifts such as geographic location, the identity of a hospital, and a temporal environment disseminated by the WILDS benchmark (Koh et al., 2021), which gives more ecologically valid evaluation of robustness when deployed in real-world environments.

The conceptualization of the common theoretical frameworks of conceptualizing distributional soundness and how it can be advanced can be covered by causal inference, domain adaptation, and worst-case optimization. Arjovsky et al. (2019) suggest learning any representations that ensure that the same predictors would appear in different environments, i.e. the representations would inherit causal features as opposed to spurious ones. Distributionally Robust Optimization (DRO) has worst-case expected loss minimization aimed at a set of distributions over which the result is distributedly robust in the sense that it is maximally insensitive to arbitrary corruptions that occur in the distribution observed upon evaluation. Group Distributionally Robust Optimization (GroupDRO, Sagawa et al., 2019) directly tries to solve the issue of performance inequality across a set of known subgroups, which has a direct relationship to the sincerity and equity of deployed systems. Most recently, pretraining foundation models on large and diverse corpora have shown greatly enhanced zero-shot distributional robustness--such as CLIP and DINOv2, for instance, have shown impressive domain robustness under adversarial perturbations and much more specialization sticking to domains unrepresented in pretraining-data--but themselves manifest characteristic failure modes under adversarial perturbation and in strictly narrow specialized domains.

### 4.3 Calibration and Uncertainty Quantification

An effective classifier must be a well-calibrated one: the probability of outcome on reaching an outcome should be related to the empirical frequency of occurrence of the outcome in examples that are presented to the classifier to compute it. Instead, modern deep neural networks are deliberately not calibrated, and they have a tendency to overconfidence, where the predicted softmax probabilities are significantly higher than the actual probabilities, which are aggravated by training on large models, long training, and a lack of explicit calibration regulation (Guo et al., 2017).

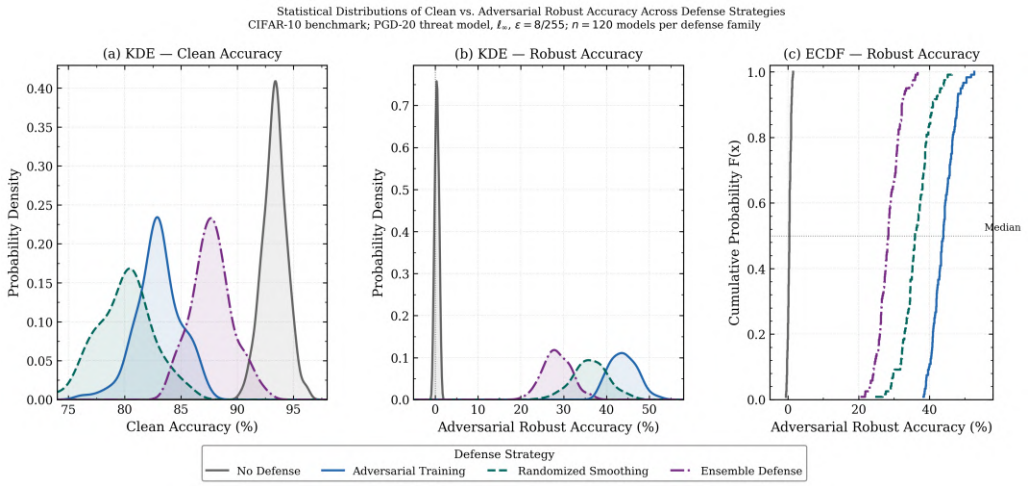


Fig. 2 Statistical Distribution of Adversarial Robustness vs. Clean Accuracy

Above Fig. 2 shows across Defense Strategies — Empirical CDF + KDE Density Curves, Shows how four principal defense families shift the joint distribution, of clean and robust accuracy on CIFAR-10 (PGD-20 threat model, ε=8/255).

The most common scalar summary of calibration quality, the Expected Calibration Error (ECE), which measures how the distribution of miscalibration is apportioned into confidence bins and how well predictions are related to accuracy in each bin, is the most commonly used. The most straightforward and commonly most effective calibration correction is temperature scaling -post-hoc division of logits by a learned scalar temperature, which can never perform any correction of positional miscalibration (in which the overconfidence is concentrated in different regions of inputs) nor can it do at all better on out-of-distribution detection.

Uncertainty quantification in deep learning aims at separating predictive uncertainty into aleatoric and epistemic parts. Aleatoric uncertainty occurs when the generating process of the data may lack an irreducible noise: e.g. ambiguity in clinical imaging because of the quality of the images, or a similar ambiguity. Epistemic uncertainty indicates the insufficiency of the model knowledge that might be filled in principle by the addition of

more data or the more detailed specification of the model. The difference is all but critical: in a healthcare diagnosis paradigm, aleatoric uncertainty ought to cause a referral to a specialist, whereas epistemic uncertainty ought to cause specific data gathering. Bayesian neural networks (BNNs) conceptualize uncertainty quantification through maintaining a posterior distribution over model parameters, in place of single value estimates but suitable approximations such as mean-field variational inference, Laplace approximations (Daxberger et al., 2021), stochastic weight averaging Gaussian (SWAG), and the more widely used Monte Carlo Dropout (Gal and Ghahramani, 2016) are all approximations of full Bayesian inference, in computational intractable deep networks. Deep Ensembles (Lakshminarayan et al., 2017) involved training many independently initialized models and combining their predictions is always superior to the uncertainty estimation of individual models and has a large baseline, but again, requires a large computational budget. The new model of Conformal Prediction (Angelopoulos and Bates, 2022) is a distribution independent, post-hoc model of building prediction sets with guaranteed coverage which only has to be given exchangeability in the underlying data and has finite sample guarantees of validity which have only recently become considered important in safety-critical applications.

### 4.4 Backdoor Attacks and Data Poisoning

Backdoor attacks can also be called Trojan attacks, and are a rather pernicious form of supply-chain attack, in which a faction corrupted a portion of training data to inject a trigger pattern: this could be an obvious patch, a frequency-domain obfuscation, or a semantic feature in conjunction with a label of misclassification. A backdoored model behaves correctly on clean inputs, and is predictable with respect to counterexample inputs, where there is a trigger in the input that causes the model to behave in a way that allows an attacker to gain control of the model behavior at inference time without raising standard accuracy measures. The original backdoor attack is BadNets (Gu et al., 2019), which implemented visible patch triggers, and the other attacks have gone on to create invisible, input-sensitive and semantic backdoors, which are much harder to detect. The threat model is especially applicable in the modern deep learning ecosystem where practitioners regularly popularize publicly released training with custom techniques, and where training-as-service services are common both of which provide adversaries with scaling.

Protections against the phenomenon of backdoor attacks cut across three paradigms: detection of contaminated samples in the training set prior to training a machine, detection of backdoor behavior in a trained machine and sanitization against backdoored models, either by fine-tuning a model or pruning it. The backdoor is identified by solving a reverse engineering problem in which the smallest universal perturbation leads to

misclassification to every target class in the model is identified (Neural Cleanse, 2019) and anomalously small perturbations are reported as the likely triggers. The inference-time backdoor input detection technique is known as STRIP (Gao et al., 2019) and it is based on the intuition that adding strong perturbations to the triggered inputs can be done without any deciphering, unlike strong perturbations to clean inputs which indeed results in a change of their classification. The latter style is spectral signatures (Tran et al., 2018), which takes advantage of the fact that the covariance spectrum of intermediate representations provides the characteristic prints of misfitting examples. This large body of defense literature notwithstanding, there is no universal detection or mitigation strategy that is applicable across all categories of attack, and attack adaptors that are sensitive to particular defenses can usually bypass them, a dynamic that is an arms race and reflects the overall problem of adversarial robustness.

### 4.5 Membership Inference, Machine Unlearning and Privacy

Deep neural networks have the capacity to store data on their training to a level where it can be used adversarially to extract sensitive data. Attacks based on membership inference (Shokri et al., 2017; Carlini et al., 2022) classify with more than a chance whether a particular example fell in a training set, thus, enabling failure in terms of health, financial or behavioral privacy when models are trained using sensitive data sets. Model inversion attacks are attacks that are able to recover representative training examples in the model parameters or query, whereas gradient inversion attacks, which is specifically applicable in federated learning, are attacks which recreate training data in gradients updates exchanged among the distributed participants. Direct extraction of verbatim training data out of the language models is the most egregious threat, as demonstrated by Carlini et al. (2021, 2023), who revealed that large language models that come with GPT-2 and other models have the ability to memorize sensitive personal information, as well as copyrighted content and training set verbatim passages.

Differential privacy (DP) gives a formal mathematical model of privacy leakage bounding measuring how varying models inputs by the inclusion or exclusion of an individual training example significantly affect model outputs by at most a sublinear (parameterized by e and d) value. Calibrated Gaussian noise to per-example gradients is also introduced to DP-SGD training to induce differential privacy (Abadi et al., 2016), but at the expense of large models, which demand large gradient noise to obtain meaningful privacy guarantees, which explains why this represents a core topic of discussion in privacy research. The new paradigm of machine unlearning includes post-deployment privacy requests, for which the model can be asked to forget certain training examples without completely retraining, motivated in part by the regulations such as the gender dataless requirements (such as the right to erasure of GDPR). Precise unlearning

needs retraining, and approximate techniques such as gradient ascent, selective fine-tuning, influence function-based removal, and memorization-targeted layer editing are computationally efficient at the unlearning performance cost area of active research as of 2024-2025.

## 5. Towards Trustworthy Deep Learning: Principles, Frameworks, and Emerging Paradigms

The reliability issues that are discussed in the list above are not isolated pathologies themselves but are symptoms of an underlying tension, namely, modern deep learning is optimized to perform averagely well on the training distribution but to operate strongly under the complexity of practical implementation, that is, to counter adversarial agents, changing distributions, and infrequent instances, and new contexts. To overcome these problems, it is necessary to have a paradigm change where reliability is seen as the post-hoc evaluation issue to a first-order design consideration that exists throughout the machine learning life cycle- both data collection and curation and architecture design, training objective specification, evaluation, deployment, and continuous monitoring.

The idea of trustful AI has been implemented in a variety of rules and technical frameworks. The AI Act (2024) is the all-encompassing binding regulatory framework of AI systems so far, by the European Union that categorises AI applications by the level of risk, and regulates high-risk systems such as technical robustness, accuracy, cybersecurity, and human oversight mechanisms. The NIST AI Risk Management Framework (2023) is a voluntary governance framework based on the functions of Map, Measure, Manage, and Govern which is relevant in the United States as it highlights the role of defining and managing uncertainty throughout the AI lifecycle. Such regulation is transforming the incentive systems of reliable deep learning research to have higher levels of formal guarantees, interpretability processes, and stringent assessment criteria that currently significantly surpass benchmark accuracy.

Mechanistic interpretability the methodical study of the computational algorithms carried out by neural network weights is one of the most attractive emergent research fields towards enhancing trustworthiness. Studies by Olah et al. (2020) and the further Circuits series of works showed that individual

bypasses and attention heads of CNNs and transformers apply understandable, interpretable curves, frequency, and syntactic structures detection algorithms, indicating that DNNs are not black box analysers, but structured information processors that can be reverse engineered. Recent studies by Anthropic researchers (Templeton et al., 2024) based on sparse autoencoders have found that of the larger models, millions of monosemantic features in large language models, and suggest that the features level of

steering and safety intervention is accessible. The factual extrapolation of these findings to the model auditing, failure mode forecasting and reliability certification is one of the key open research agendas.

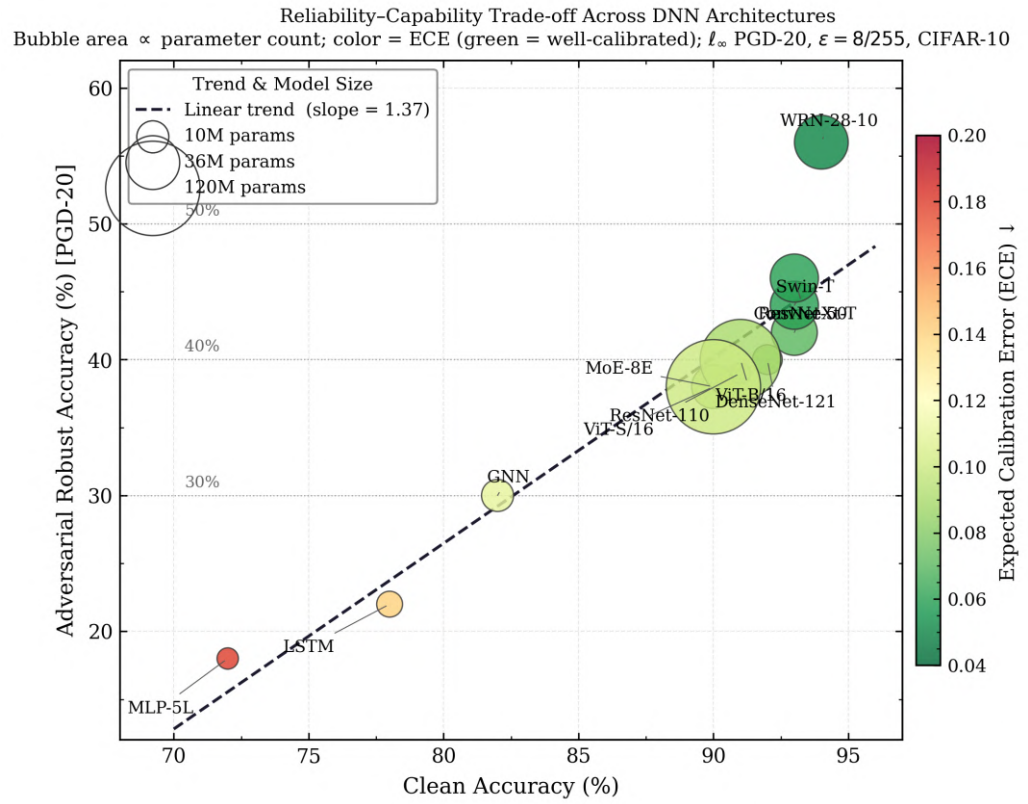


Fig.3 Reliability–Capability Trade-off Bubble Chart

Above Fig.2 is 2D scatter of Clean Accuracy vs. Adversarial Robustness for 12 architectures, with bubble area ∝ model parameter count and color encoding calibration ECE. Annotated trend lines reveal the empirical robustness–accuracy trade-off.

The new reliability considerations and capabilities of foundation models and their adaptation through parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA), prompt tuning, adapter layers, and prefix tuning, have been presented. The process of fine-tuning a large pretrained model on a downstream task can bring both task-specific threats and some behaviours cherished by the task, and deep forgetting of the more general abilities. Efforts to continue to learn an existing knowledge in sequential fine-tuning methods such as Elastic Weight Consolidation, Progressive Neural Networks, as well as replay-based methods have aimed to preserve existing knowledge but their effectiveness to scale to models with billion of parameters is an active field of research. Through the safety alignment of large language models

both using Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI (Anthropic, 2022), training objectives have been shown to be manipulable to display desired properties, though alignment brittle natureswhere allowed models behave poorly under adversarial prompting or distribution shift are an issue of serious concern as has been well-documented in the research on red-teaming.

Summary Tables

Table 1: Comparative Study of the Deep Neural Network Construction and Reliability Profile.

The next table presents a tabular comparative summary of the main families of deep neural network architecture referred in this chapter along with their main computational principles, depth profiles, and the main vulnerabilities to their reliability, relevance of current texts published between 2023 and 2025. This comparison shows that no given architecture can be robust in all dimensions of reliability and this has spurred the creation of hybrid architectures and reliability-conscious architectures as the future research frontier.

| **Architecture** | **Core Mechanism** | **Depth Profile** | **Key Reliability Issues** | **Emerging Variants (2023–2025)** |
|---|---|---|---|---|
| Feedforward MLP | Fully-connected linear layers with nonlinear activations | Shallow to moderate; vanishing gradients limit depth | Overconfident predictions; poor OOD detection; susceptibility to gradient-based attacks | Kolmogorov-Arnold Networks (KANs); Neural Ordinary Differential Equations |
| Convolutional Neural Network (CNN) | Hierarchical spatial feature extraction via shared filters | Very deep (e.g., ResNet-1000+); skip connections mitigate vanishing gradients | Texture bias; adversarial patch vulnerability; distribution shift brittleness | ConvNeXt V2; DenseNet hybrids; Depthwise Separable Variants |
| Recurrent Neural Network (RNN/LSTM) | Sequential hidden state propagation with gated memory | Temporal depth bounded by sequence length; exploding gradient risk | Long-range dependency failure; temporal adversarial perturbations; compounding uncertainty | Mamba (State Space Models); xLSTM; Griffin architectures |
| Transformer (Self-Attention) | Scaled dot-product attention with positional encoding | Deep stacks (e.g., 96-layer GPT-4); quadratic attention complexity | Hallucination; sycophancy; prompt injection; calibration failure at scale | Flash Attention 3; Sparse Transformers; Mixture-of-Experts (MoE) |
| Graph Neural Network (GNN) | Message-passing aggregation over non-Euclidean graph structure | Limited effective depth due to over-smoothing | Over-squashing; structural adversarial attacks; heterophily failure | GraphTransformer; MPNN++; Directional GNNs |

| | | | | |
|---|---|---|---|---|
| Capsule Network | Dynamic routing between capsule units encoding part-whole relationships | Moderate; routing iterations add computational depth | Scalability constraints; limited robustness gains in practice; small-sample sensitivity | E-CapsNet; Efficient Capsule Routing with Attention |
| Variational Autoencoder (VAE) | Latent space regularization via KL divergence with Gaussian prior | Encoder-decoder symmetric; bottleneck controls expressivity | Posterior collapse; blurry reconstruction; latent space misalignment under covariate shift | Hierarchical VAE; VQVAE-2; Diffusion-VAE hybrids |
| Generative Adversarial Network (GAN) | Minimax game between generator and discriminator networks | Generator and discriminator each deep; training dynamics fragile | Mode collapse; training instability; memorization of training data; deepfake risks | StyleGAN3; Projected GAN; Consistency Models |
| Diffusion Model | Iterative denoising via learned score functions or noise prediction | U-Net backbone; depth determined by denoising steps | Slow inference; watermarking challenges; adversarial conditioning; content poisoning | DDIM; Latent Diffusion; Flow Matching; Rectified Flow |
| Mixture-of-Experts (MoE) | Sparse gating routes inputs to specialized sub-networks | Very deep with selective activation; efficient at scale | Load imbalance; expert collapse; routing instability; interpretability deficits | Switch Transformer; Mixtral-8x22B; DeepSeek-V2 MoE |

Table 2: Taxonomy of Reliability Failure Modes in Deep Learning Systems

The following table taxonomizes the major reliability failure modes affecting deployed deep learning systems, providing for each dimension a precise problem definition, representative instantiations in the attack or failure mode literature, current state-of-the-art detection and mitigation strategies, and the benchmark protocols used for standardized evaluation. This taxonomy reveals the breadth of the trustworthy deep learning research agenda and the degree to which different failure modes require distinct conceptual frameworks and technical approaches.

| **Reliability Dimension** | **Core Problem Definition** | **Representative Attack/Failure Mode** | **Detection / Mitigation Strategy** | **Benchmark / Evaluation Protocol** |
|---|---|---|---|---|
| Adversarial Robustness | Model misclassification under imperceptible input perturbations | PGD (Projected Gradient Descent) attacks; AutoAttack | Adversarial training; certified defenses (Lipschitz bounds, randomized smoothing) | RobustBench; AutoAttack leaderboard; CIFAR-10-C |
| Distributional Robustness | Performance degradation | Texture shift; corrupted benchmarks | Data augmentation; domain adaptation; | ImageNet-C; WILDS benchmark; |

| | | | | |
|---|---|---|---|---|
| | under covariate or concept shift | (ImageNet-C); domain gap | invariant risk minimization (IRM) | DomainBed; BREEDS |
| Calibration Failure | Mismatch between predicted confidence and empirical accuracy | Overconfident softmax outputs; ECE exceeding acceptable thresholds | Temperature scaling; isotonic regression; label smoothing; Dirichlet calibration | Expected Calibration Error (ECE); Reliability diagrams; MCE |
| Uncertainty Quantification | Inability to distinguish aleatoric from epistemic uncertainty | Silent failure in OOD regions; confident wrong predictions | Monte Carlo Dropout; Deep Ensembles; Bayesian neural networks; ConformalPrediction | AUROC on OOD; Uncertainty Benchmarks; OpenOOD; AnomalyCLIP |
| Backdoor / Trojan Attacks | Hidden functionality triggered by specific input patterns | BadNets; blend-based triggers; invisible frequency-domain backdoors | Neural Cleanse; STRIP; spectral signatures; meta-neural analysis | TrojAI benchmark; BackdoorBench; BAARD evaluation suite |
| Data Poisoning | Corruption of training data to degrade or redirect model behavior | Gradient-matching attacks; bullseye polytope; clean-label poisoning | Data sanitization; certified learning; influence function–based filtering | CleanLab; SNAP; Poisoning Benchmark (Witches Brew) |
| Privacy Leakage | Extraction of sensitive training data via model queries | Membership inference; model inversion; gradient inversion attacks | Differential privacy (DP-SGD); federated learning; machine unlearning | MIA benchmarks; PrivacyRaven; ML Privacy Meter |
| Shortcut Learning | Reliance on spurious correlations instead of causal features | Background bias; watermark-label correlation; gender stereotyping in NLP | Causal representation learning; debiasing via reweighting; counterfactual augmentation | Waterbirds; CelebA-bias; NLI challenge sets; MNLI-m |
| Catastrophic Forgetting | Loss of prior task performance during sequential learning | Accuracy degradation on Task A after fine-tuning on Task B | Elastic weight consolidation (EWC); progressive networks; rehearsal methods; LoRA continual | Split-CIFAR; Permuted MNIST; CORe50; CLiMB benchmark |
| Interpretability Deficit | Opaque decision-making processes preventing auditing or trust | Saliency attribution disagreement; explanation faithfulness failure | SHAP; Integrated Gradients; concept-based explanations; mechanistic interpretability | Faithfulness metrics; ROAR; ERASER; AI explanations benchmark |
| Model Collapse | Degradation of generative models trained on AI-generated data | Quality and diversity decay in iterative synthetic data pipelines | Human-data anchoring; curated data curation pipelines; provenance tracking | FID/IS degradation curves; model collapse benchmarks (Shumailov et al., 2024) |

## 6. Conclusion

The chapter has laid the mathematical and architectural make-up of the deep neural networks and systemically enumerated the multidimensional reliability issues that will have to be resolved before the systems can be flattered to be reliable enough to be deployed consequentially. The capability-reliability mismatch is not an accident but instead a structural feature: the very features which allow deep networks to learn rich and high-dimensional representations, namely non-convex optimization surfaces, large number of parameters, implicit regularization, and emergent compositional representations, also lead to their typical vulnerabilities of adversarial sensitivity, distributional brittleness, miscalibration, and privacy leakage. The problems of these vulnerabilities can be addressed by a set of theoretical understanding, innovative architectural design, redesign of training goals, post-hoc methods of calibration and detection, and governance systems that make models occur within their overarching sociotechnical deployment environment.

A number of cross-cutting themes are taken out of the reliability literature that will be evolved in details in later sections of the chapters. To begin with, the equilibrium between attacks and defenses, which has been the most evident in the literature of adversarial robustness, is indicating that very little about defenses is stable, and that robustness itself should be measured through adaptive adversaries. Second, the quality and representativeness of the training data can not be decoupled and the reliability of a deployed model, as it is not a folk law but rather a mathematically proven fact in the data poisoning and shortcut learning literature. Third, both directions make the reliability situation more complex because extremely large-scale foundation models have recently demonstrated remarkable emergent robustness behavior on several dimensions (especially distributional robustness properties in various dimensions) and also prominently new and poorly comprehended failure modes (especially hallucination, sycophancy, and prompt injection within the language domain). Fourth, regulatory systems are swiftly changing, and are beginning to introduce binding mandates of robustness assessment, documents of uncertainty, and human control with which the technical community must proactively interact with.

Trustworthy deep learning has been a fruitful area of research undergoing an unprecedented frenzy with the core contributions being made at an increasing pace in conflicting areas such as adversarial robustness, uncertainty quantification, privacy-preserving learning, interpretability and alignment. This chapter has given both the conceptual vocabulary and technical stands of the following chapters, which study a particular dimension of trustworthiness with more technical ingredients. The general concept of this book is that trustworthiness stands not as a marginal addition to the power of deep learning systems and is its necessary complement, otherwise the use of powerful deep learning systems in high-stakes systems is an unacceptable risk to society but with

it such systems have the power to produce genuine benefits across medicine, science, engineering and human welfare.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 3: Uncertainty in Deep Learning: Aleatoric and Epistemic Perspectives

## Abstract

Deep learning systems have shown a tremendous predictive accuracy on a wide range of applications, but their use in high-stake fields still begs deeper questions as to the reliability of their results. At the centre of such a challenge of trustworthiness lies the quantitative and communicative expression of uncertainty on the basis of principles. The chapter critically and thoroughly discusses uncertainty in deep learning distinguishing between aleatoric uncertainty, irreducible noise and ambiguity in the data-generating process, and epistemic uncertainty, due to a lack of knowledge and limited training data, which in theory can be reduced with additional information. We investigate the theoretical framework of both types of uncertainties, overview the significant methodological paradigms that have been designed to approximate, perform decomposition of, and communicate these quantities, and also discuss the theoretical progressions in detail such as evidential deep learning, conformal prediction, post-hoc calibration methods, and uncertainty-aware large language models. Special focus is on realistic deployment issues, the theory of calibration, and future issues at the boundary, in both generative models and safety critical systems, of trying to estimate uncertainty when distribution shifts. Two broadly reference tables are used to synergize methodological and application-domain landscapes to give researchers and practitioners some action orientation.

## 1. Introduction

Application of deep learning models to consequential real-world problems, such as clinical diagnosis to autonomous navigation, financial forecasting to scientific discovery, and so on, require not only high average-case performance but also good self-awareness of the boundaries of the corresponding performance. A model that is certain when uncertain, or diffuse when certain, places failure modes of qualitatively different

character on the systems in which it occurs. This appreciation has triggered a tidal wave of academic and engineering attention to uncertainty quantification (UQ) as a first-class design issue in deep learning designs, training algorithms, as well as training evaluation strategies. Nonetheless, even in the light of the maturity of the profession underlying conceptual differences are understated in practice with the effects on the perception, implementation, and reliance on uncertainty estimates.

The first underlying difference in the literature on uncertainty is that between aleatoric and epistemic uncertainty, dating back to the epistemological and decision-theoretical literature, and then transferred to the statistical and machine-learning literature. Aleatoric uncertaintyAlea, as Latin alea, denotes the intrinsic randomness, which cannot be reduced, of the generating mechanism of the data. Noise in a medical scanner, the inherent ambiguity of the border between two types of tissue, stochastic results of a quantum physical process, etc, all of these are sources of aleatoric uncertainty that cannot be removed by any extra information or any more expressive model. Epistemic uncertainty, which is a Greek word episteme, knowledge (ensured) indicates what the model does not know as the result of incomplete coverage of the input space, sparse training data or misspecification of the model family. In contrast to its aleatoric counterpart, the uncertainty of epistemics is, in theory, reducible: an expert that gets additional more informative training cases, increases the hypothesis space of the model or adds prior knowledge of the domain can predict that it will decrease.

This difference has far-reaching consequences on the design of the system. The preferable measure of epistemic uncertainty is epistemic uncertainty in safety-critical applications: it indicates areas of input space in which model predictions are to be treated with caution and where human supervision and extra data gathering is likely to be the most justified. Instead, aleatoric uncertainty tells the communicator about the inevitable irreducible periods of prediction to downstream decision-makers, and influences even the design of loss functions to not punish the model because of inherently noisy targets. When it becomes uncertain due to a pathology seen by a clinical decision support system has never been seen before (epistemic) versus when there is fundamental disagreement over a specific diagnosis even with the presence of the entire picture (aleatoric), it should act differently. These two sources are easily mixed and may result in ill-calibrated measures and, in the end, failures of credibility.

This chapter is structured in such a way that it follows a path of theoretical basis up to methodological innovations to the new application boundaries. In section 2, the formal probabilistic framework is determined. In Section 3 the aleatoric uncertainty is explored, including heteroscedastic models, distributional regression and the label noise. Section 4 shifts to epistemic uncertainty, where it focuses on Bayesian neural networks, approximate inference and ensemble-based methods. Section 5 deals with the breaking down of total predictive uncertainty and calibration of uncertainty estimates. Section 6

presents new frontiers such as evidential deep learning, conformal prediction and uncertainty in LLM. Section 7 addresses the fields of application and real implementation. Section 8 gives 2 integrative reference tables. Section 9 points out the open research frontiers and Section 10 comes to an end.

2. Such Probabilistic Foundations of Uncertainty in Predictive Models.

To make formal arguments surrounding uncertainty in deep learning, it would be necessary to base the argument on a rational probabilistic foundation. Where x is an input-vector, and y a target variable, and we get access to a training dataset D = {(x_i, y_i)}_{i=1}^{N} which are chosen independently according to an unknown data-generating distribution p*(x, y). The objective in the typical supervised learning paradigm is to have a parametric deep neural network that approximates a conditional distribution theta* (y x, theta) parameterized by weights theta. Prediction using a single conditioned model theta (a point estimate, either the maximum a posterior estimate or the maximum likelihood estimate) confounds two terms of variation: the randomness of p*(y | x) itself when this is done at the true parameter value, and the uncertainty of theta itself due to the finite size of D and the constraint of the optimization surface.

The Bayesian framework provides the natural language for disentangling these contributions. Within this framework, uncertainty about the parameters is encoded in a posterior distribution p(theta | D) obtained by updating a prior p(theta) with the likelihood of the observed data. Predictions are then made by marginalizing over the posterior, yielding the posterior predictive distribution p(y | x, D) = integral p(y | x, theta) p (theta|D) d(theta). The variance of this distribution decomposes, under mild assumptions, into an aleatoric component—the expected conditional variance and an epistemic component—the variance of the conditional mean. This decomposition, made precise through the law of total variance, provides the mathematical scaffolding upon which most practical UQ methods are constructed, even when full Bayesian inference is computationally intractable.

The computational complexity of computing p(theta | D) directly due to the large dimensionality and non-convexity of the parameter space makes it hard to estimate the probability distribution of deep neural networks, which in turn has led to the invention of a diverse array of approximate inference algorithms. Variational inference replaces the intractable posterior using a parametric family and maximizes the evidence lower bound (ELBO), and is basically minimizing the Kullback-Leibler divergence between the variational approximation and the true posterior. Hamilton Monte Carlo (and modern stochastic gradient variants of it, referred to as SGLD and SGHMC) is a type of Markov chain Monte Carlo that is asymptotically accurate at high probability of sampling, but at a high computational cost. Laplace approximations are based on a second-order Taylor expansion of a point estimate to build a surrogate of the posterior that is a Gaussian. All

methods make certain assumptions and trade offs in their computation that have the implications to the fidelity of uncertainty estimates whose implications have been extensively the subject of empirical research.

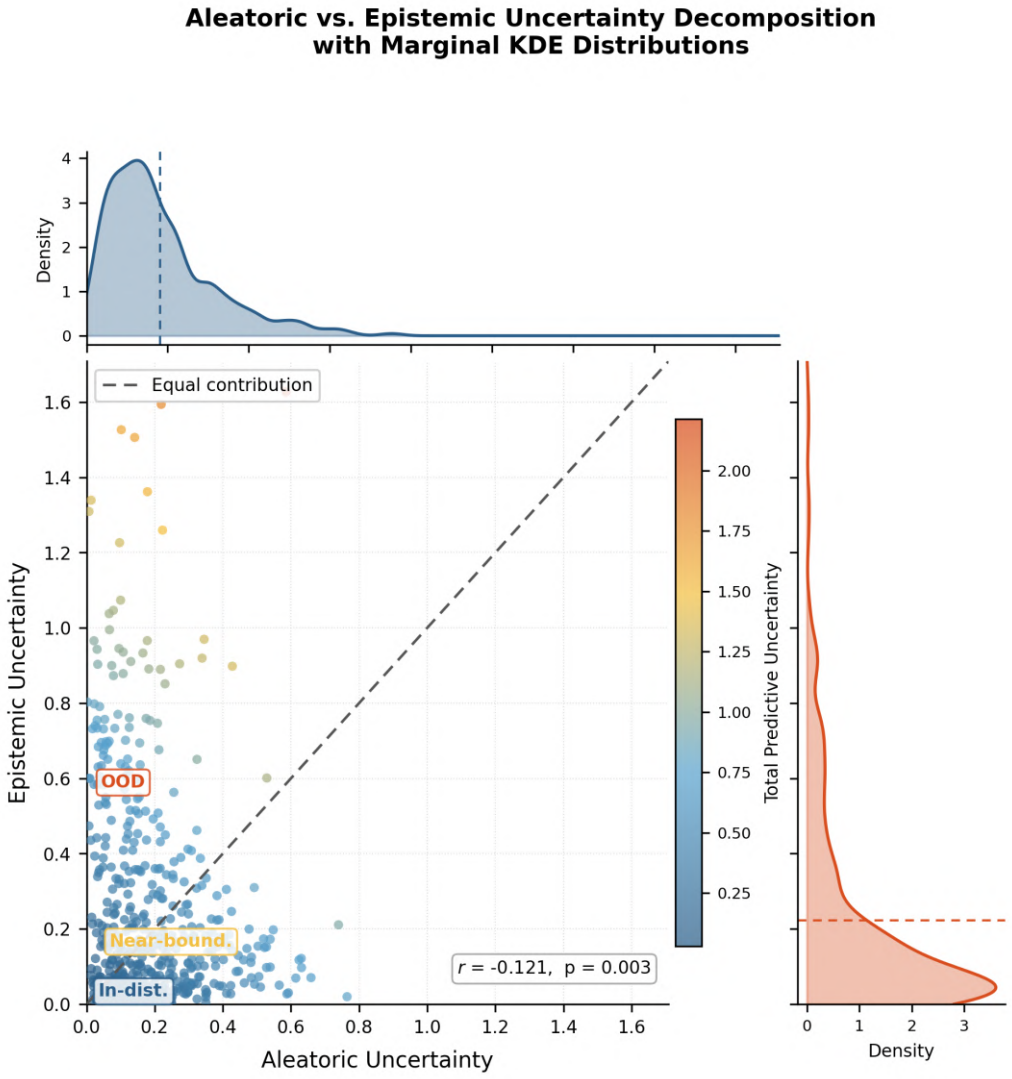


Fig.1 Aleatoric vs. Epistemic Uncertainty Decomposition with Marginal KDEs

Fig 1 shows A joint scatter plot mapping each test sample's aleatoric uncertainty (x-axis) against its epistemic uncertainty (y-axis), with point colour encoding total predictive entropy. Three annotated clusters — In-distribution, Near-boundary, and Out-of-distribution — illustrate how the two uncertainty types co-vary across input regimes. Marginal KDE density curves are embedded along each axis. A dashed diagonal marks equal aleatoric/epistemic contribution, and Pearson *r* is annotated.

An important notion, on which the practical assessment of UQ methods is based, is calibration. It can be said that a probabilistic predictor is calibrated when the confidence levels that it expresses are equal to the empirical frequencies: on all occasions when the model allocates probability p to any of the binary events, a fraction p of these occasions should actually result in the positive outcome. The assessment of calibration may be done using reliability diagrams and summarized as the Expected calibration error (ECE) or Maximum Calibration error (MCE). The most widespread failure mode in the modern deep networks is overconfidence, in which predicted probabilities of the network are consistently more accurate than the accuracy in the empirical performance of the net in the confidence bins. The other phenomenon namely underconfidence is also less prevalent but equally negative in areas where bold suggestions are needed. Post-hoc recalibration methods that are most commonly applied include temperature scaling, Platt scaling and isotonic regression which modulates the predicted probabilities without re-training the network.

## 3. Aleatoric Uncertainty: Modelling, Estimation, and Sources.

### 3.1 Taxonomy of Aleatoric Uncertainty

The deep learning contexts with aleatoric uncertainty have a number of different mechanisms that necessitate the different modeling strategies. Homoscedastic aleatoric uncertainty is the assumption that the level of noise is the same throughout the input space, to all predictions On average, the same amount of inherent randomness applies, independent of which pattern of input features takes place. Although have the convenience of analytical ease, this assumption is hardly ever justified in practice. Much more widespread, and much more significant is heteroscedastic aleatoric uncertainty, in which the amount of noise is determined by the input: a natural image with motion blur has even more irremovable perceptual ambiguity than a perfectly focused one; particularly in a medical scan, patient-independent noise in the signal that cannot be eliminated with any further training data is introduced. Heteroscedastic modeling is thus not just an important statistical nicety but for reliable application in most practical applications.

Another difference is that of continuous and categorical output settings. Aleatoric uncertainty is a natural form of uncertainty in regression tasks, where the variance or complete distributive characteristics of the conditional output distribution p(y x, theta) can be trained in the network to predict in either regression tasks with this form of uncertainty. During classification anymore classification, the softmax values are other times confused with aleatoric uncertainty, yet this description is flawed: a big-entropy

softmax sample might be both representative of increased label ambiguity (aleatoric), and signals that the sample is not in-distribution (epistemic) by the model. In the classification context, to eliminate these two without coming up with an architecture would either involve designing higher-order forms of uncertainty like the Dirichlet distortion over the space of simplex class frequencies.

### 3.2 Heteroscedastic Neural Networks

The heteroscedastic neural network, which was introduced in influential work by Kendall and Gal (2017) as the canonical method of estimating the input-dependent aleatoric uncertainty in regression and was subsequently formalized in the literature, is the canonical method of estimating the input-dependent aleatoric uncertainty in regression. The network has two output heads, at this model the predictive mean mu(x) and the log-variance log(sigma 2 (x )) in which sigma 2 (x ) is the level of aleatoric noise at input x. Training the network is performed by minimizing the negative log-likelihood using an observation model that is based on a Gaussian, and this is equivalent to a loss-function that not only consists of prediction error terms multiplied by the predicted variance, but also regulator terms representing the log-determinant of the estimates of the variance themselves. The advantage of this formulation is that it automatically learns to place more uncertainty on inputs whose targets have noisier values, without the need to have noise labels in the training data.

More generalizations of the heteroscedastic model have been introduced to deal with non-Gaussian models of observation, multi-output models, and also when dealing with structured outputs. In the case where the noise distribution is heavy tailed Student t-likelihood is more robust to outliers whilst retaining the ability to compute a closed form gradient. Those arbitrary complex conditional distributions p(y  x) that are needed to model multi-modality and skewness of the aleatoric distribution and are otherwise unattainable by Gaussian models can be modeled using normalizing flows. Per pixel heteroscedastic uncertainty maps have been found diagnostically useful in semantic segmentation tasks, with uncertainty focused at object boundaries as well as where annotation is conflicting across learners- a observation that validates the interpretability of the learnt uncertainty representation as well as, indicating potential use as a probabilistic guide to active learning and curriculum choices.

### 3.3 Aleatoric Uncertainty and Noise on labels.

There is a close relation between aleatoric uncertainty and label noise, an aspect that should not be overlooked. The disagreement between annotations is a direct expression of aleatoric ambiguity as to which target the annotations actually are when these labels

are produced by multiple annotators whose expertise or interpretive structure is different. When one learns using such data without explicitly modeling this ambiguity the result of the learning instead of the latter, p(y  x), is the learning of an average over the annotator distribution. Recent challenges in learning through crowds, and multi-annotator networks have attempted to fix this through learning a model of the behaviour of annotators and the predictive model itself in co-operation to recover a noisier model of the aleatoric signal and a fatter model of human uncertainty that can be delivered to end users.

The connection between aleatoric uncertainty and adversarial inputs is also an issue that should be closely regarded. Contrived adversarial examples that are close to the fallacies of natural decision-boundaries may seem, on a pure aleatoric view, to be representatively ambiguous, when, in reality it is adversarial manipulation that causes the ambiguity rather than natural variability in data. This fact leads to the design of uncertainty estimators resistant to adversarial perturbations--a problem that relates the UQ literature to wider problems of adversarial robustness elsewhere in the rest of this book. Multiple more recent suggestions have integrated adversarial training together with objective measures of uncertainty to generate models with estimates of uncertainty simultaneously more well-calibrated and robust to targeted perturbations on the input.

## 4. Epistemic Uncertainty: Approximations, Methods and Frontiers.

### 4.1 Bayesian Neural Networks and The posterior Inference Problem.

The ethical treatment of epistemic uncertainty in neural networks will result in the Bayesian neural network (BNN) with the model weights being modeled as random, with a posterior distribution given the observed data. The allure of the BNNs is immense: through the explicit representation of the uncertainty about model parameters, they also offer a mathematically-motivated way to express what the model knows and what it does not know, and also automatically generate predictive distributions instead of point predictions. Nonetheless, any exact Bayesian inference in BNNs is computationally infeasible, except in the case of the smallest networks, because it involves integrating on an a posteriori distribution that exists in a non-convex high-dimensional space given by the compound volition loss scene of deep networks.

Similar to what an actual Bayes by Backprop algorithm has demonstrated, mean-field variational inference (Blundell et al., 2015) uses a fully factored Gaussian approximation to the posterior and optimizes the ELBO using reparameterized gradients. Being computationally manageable, the mean-field assumption provides a structural bias that tends to underestimate posterior covariance and in practice fails to provide accurate

representations of uncertainty. Variational families Structured variational families, such as matrix-variate Gaussians, normalizing flow posteriors or hierarchical variational models, partially solve this restriction at the expense of much more implementation complexity and computational cost. Stochastic Gradient Langevin Dynamics (SGLD) and extensions based on MCMC techniques do not face the mean-field bias; however, they also are susceptible to hyperparameter choices and take students of time, preventing their use with large-scale architectures without substantial engineering resources.

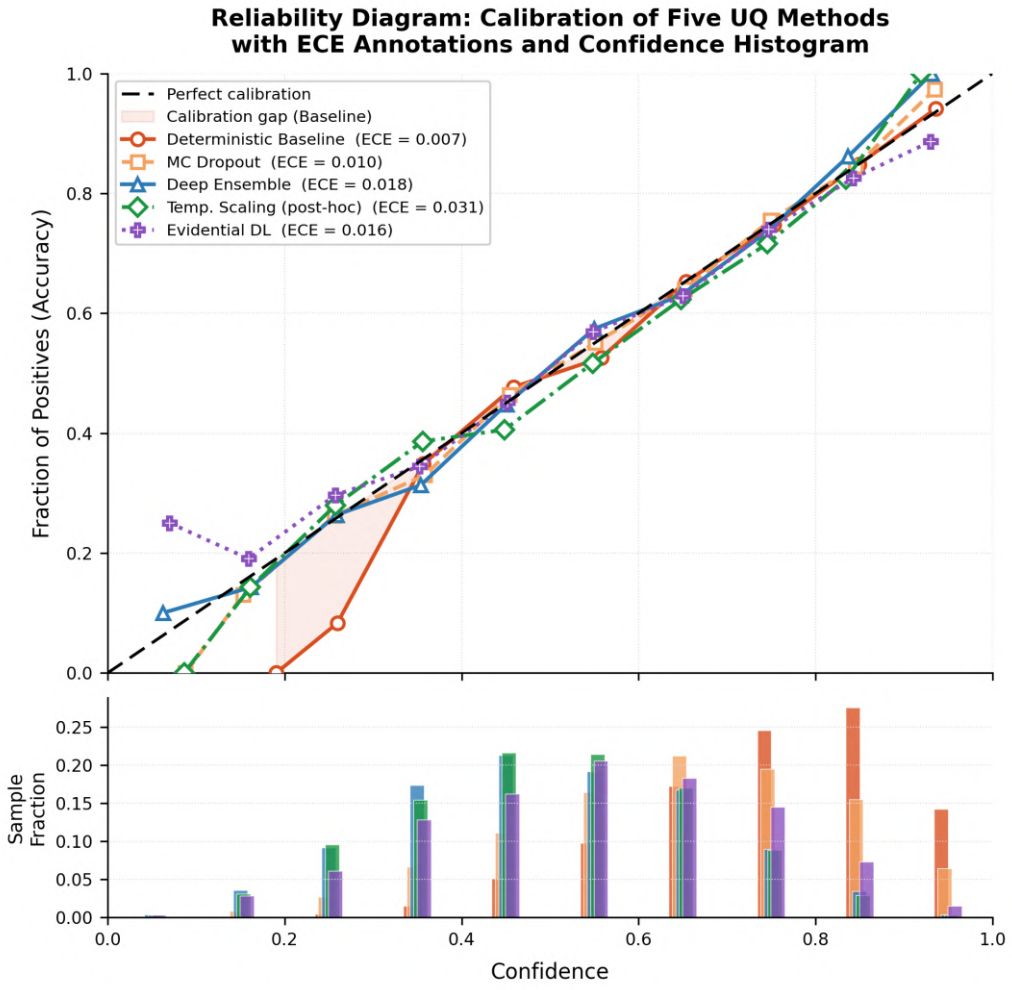


Fig.2 Calibration Reliability Diagram with ECE Annotations and Confidence Histogram

Above fig. 2 shows A two-panel reliability diagram comparing five canonical UQ methods (Deterministic Baseline, MC Dropout, Deep Ensemble, Temperature Scaling, Evidential DL). The upper panel plots confidence bins against empirical accuracy; the diagonal represents perfect calibration, and ECE values are embedded in the legend. A shaded region highlights the calibration gap for the worst-performing method. The lower

panel shows the confidence histogram (sample fraction per bin) for all five methods, revealing over-confidence patterns and the effect of post-hoc recalibration.

### 4.2 Deep Ensembles and Ensemble-Based Epistemic Uncertainty.

One of the most successful and reproducible methods to estimate epistemic uncertainty in deep learning has been proposed by Lakshminarayanan et al. (2017), based on deep ensembles which have been in turn developed as an alternative to Bayesian methods that can be considered practical in many applications. In an ensemble, there are M randomly abstracted neural networks and this network is trained on the same dataset and the prediction is collected through averaging of predictive distributions. The variety of ensemble members triggered by the stochastic gradient optimization algorithm and unstructured weight introduction is used as an approximation to sampling of different modes of the posterior. In practice, ensembles have been shown to solve calibration tasks, out-of-distribution detection problems, and distributional robustness tests in practice, setting a very high practical standards of any suggested UQ methodology.

The theoretical insight into the reasons of the effectiveness of the work of ensembles has been enriched in recent years. The relationship between ensemble diversity and the epistemic uncertainty decomposition has been formalized in terms of bias-variance tradeoff and by information-theoretic analysis of mutual information between predictions and parameter as a natural epoch of the epistemic uncertainty is apparent with the disagreement between group members. More efficient ensemble construction methods have also been investigated recently, such as Batch Ensembles (where most parameters are shared among ensemble members), MIMO architectures (where different inputs are processed by a single forward pass), and Hyper-Ensembles which are constructed by exploiting learned diversity, instead of random initialisation. Such methods aim to maintain the epistemic coverage of ensembles but there is a drastic decrease in the amount of computation and memory required to represent them.

### 4.3 Video Monte Carlo Dropout and Lightweight Bayesian Approximations.

Although initially viewed as a regularization method, Gal and Ghahramani (2016) have refined dropout to be an approximation of a Bayesian inference method. With dropout masks kept on during inference, and with T stochastic forward passes, it is possible to obtain a sample of an approximate posterior predictive distribution without updating the training procedure. The variance between T predictions is then used to estimate the epistemic uncertainty and the average give the point estimate. What is elegant about MC

Dropout is that it can be applied to any dropout-regularized network, but its weakness is similarly well-documented: quality of the Bayesian approximation can be adversely affected by the dropout probability and the architecture, and the procedure can also give overconfident uncertainty estimates in areas that are far away.

Even more recent lightweight approximations are Stochastic Weight Averaging Gaussian (SWAG), that trains a Gaussian approximation to the posterior by aggregating statistics on weight-based metrics using a stochastic gradient descent history of cyclic learning rates; and the Laplace approximation of the last or full network, that trains a Gaussian approximation to the posterior by computing/approximating the Hessian of the negative log-likelihood at the MAP estimate. The similarity between these approaches is that they aim to get simple Bayesian uncertainty approximations at comparatively low computation cost, and thus they are appealing to practitioners who cannot afford the overhead of ensembles or MCMC sampling. Such methods have been compared in a systematic basis of large scale benchmarking experiments, with the best known being the Uncertainty Baselines benchmark by Google Brain and the PLEX framework demonstrating that none of these methods performs better in all evaluation criteria.

## 5. Breaking down and Calibrating Total Predictive Uncertainty.

The fact that it can break down total predictive uncertainty into its aleatoric and epistemic components is not just its theoretically desirable property, but it also has direct practical implications in the application of uncertainty information in downstream decision making. The decomposition under the law of total variance is easy to apply in any ensemble-based or Bayesian model in regression problems: the expected value of the individual predictions of the per-member predictions represents aleatoric uncertainty, and the anticipated value of the individual means of the per-member predictions represents epistemic uncertainty. In the case of classification, the associated decomposition is finer grained, it is usually done in information-theoretic terms: the total uncertainty which is the entropy of the predictive distribution, the epistemic uncertainty which is the mutual information between predictions and parameters (also called BALD-Bayesian Active Learning by Disagreement), and the aleatoric uncertainty which is the expected entropy of the discrete model predictions.

The calibration examination extends on bilateral statements of either being overconfident or underconfident to the entire reliability of the predictive distribution. The common measure of evaluation of the classification UQ literature has been Expected Calibration Error (ECE) which has been shown to be sensitive to binning strategy and does not provide insights into the miscalibration conditioned by classes so has been supplanted by more detailed calibration diagnostics such as the Integrated Calibration Index, kernel-based calibration tests, and more recently token-conditioned

miscalibration measures of language generation models. To get distributional accuracy measures beyond simple mean squared error, which are required in regression, the Continuous Ranked Probability Score (CRPS) and interval coverage rates are used. Appropriate scoring criteria: loss functions which are minimized by the predicted distribution being the actual distribution whose distribution can construct a theoretical framework to design training goals which reward the expression of uncertainty in a calibrated way.

Post-hoc Although there are practical paths to post-training calibration without adjusting the model, post-hoc methods can help strategies that lead to better calibration. Temperature scaling, which is the least difficult and frequently most useful, also only learns one scalar which is used to rescaling the logits of a softmax classifier to reduce ECE on a held-out validation set. This is generalized in Dirichlet calibration to a multi-class context which includes class-conditioned corrections. Platt scaling and Isotonic regression are non-parametric and parametric alternatives respectively with various bias-variance tradeoffs. A key caution to all post-hoc approaches is that they can do a better job of marginal calibration which is being averaged over the test distribution, but can possibly not fix any conditional miscalibration which might be present, especially in relation to subgroups or regions of input that are underrepresented in the calibration validation set. This weakness has provided the incentive to develop class-conditional and input-conditional calibration techniques which provide more robust guarantees in structured contexts.

## 6. New Paradigms in the Uncertainty Quantification.

### 6.1 Evidential Deep Learning

The new paradigm of uncertainty quantification that is completely independent of multiple forward passes and ensembles is called evidential deep learning (EDL) and proposed by Sensoy et al. (2018) to achieve classification and later extended to regression by Amini et al. (2020). Its fundamental principle is based on Dempster-Shafer theory of evidence and subjective logic: instead of making a prediction of the initial level such as a categorical distribution, or a Gaussian prediction, the network provides the prediction of the parameters of a more confusion number distribution, like a Dirichlet distribution over the class probability simplex (used with classification) or a Normal-Inverse-Gamma distribution over the mean-variance couple (used with regression). Such higher-order distributions represent aleatoric uncertainty (in the sense of the dispersion of the underlying base distribution which they model) and epistemic uncertainty (in the sense of the total evidence or concentration of the higher-order distribution).

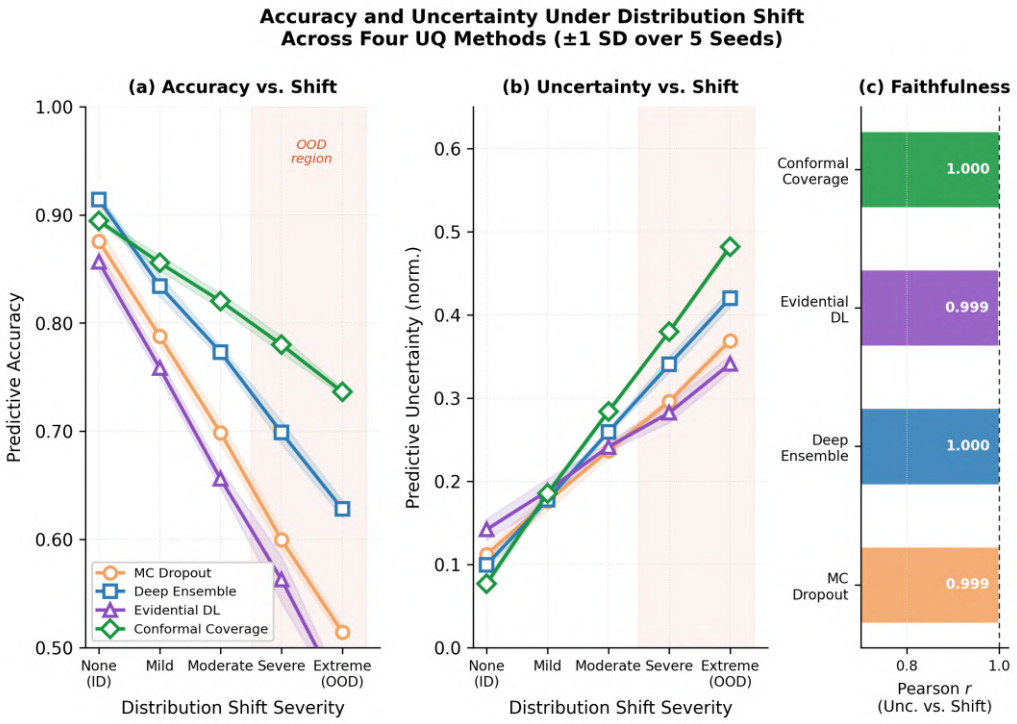


Fig.3 Accuracy and Uncertainty Under Distribution Shift (Three-Panel)

Fig. 3 visualised A three-panel figure showing (a) accuracy decline and (b) uncertainty increase across five levels of distribution shift severity (None → Extreme OOD) for four UQ methods, with ±1 SD shaded bands over five random seeds. Panel (c) ranks methods by Pearson r between uncertainty and shift severity — a direct measure of uncertainty faithfulness as defined by Ovadia et al. (2019). The OOD region is shaded red in both line-plot panels.

EDL has the strengths that the method is computationally efficient, a single forward run is sufficient to compute distinct aleatoric and epistemic uncertainty estimates, and it is principled, based on evidence theory. Nevertheless, the model has come under great criticism. Some of them have found pathological behaviors such the model assigning high vacuity (low evidence) to in-distribution examples and low vacuity to some out-of-distribution inputs, especially those in areas of high-density under the estimated evidence distribution. New theoretical studies have cast doubt on whether EDL is really the expression of Bayesian epistemic uncertainty or is a heuristic that may perform well on some criteria but do not have a strong role in the sense of a probabilistic basis. Continuous efforts are under way to overcome these drawbacks with altered training

goals and supplementary OOD exposures as well as hybrid methods using evidential priors and ensemble or Bayesian elements.

### 6.2 Conformal Prediction

Conformal prediction has become one of the most conceptually developed and empirically convincing models of uncertainty quantification, with coverage guarantees being finite-sample, and being completely distribution-free, assuming exchangeability. Conformal prediction, which works by wrapping any underlying deterministic or probabilistic base model, unlike Bayesian methods which require the specification of priors and approximate inference algorithms, is a conformal prediction projection that produces either prediction sets (in the case of classification) or predictions intervals (in the case of regression) that are known to contain the true label with at least probability 1 - alpha of the probability level chosen by the user. This conclusion is independent of the underlying data distribution, model structure or even sample size, and conformal prediction is the only method of prediction with some desirable property that finds application at high stakes, where worst-case coverage of prediction is more valued than average-case calibration.

Recent trends have significantly increased the scale and the accuracy of conformal prediction of deep learning. Split conformal prediction, where the nonconformity scoring is computed with the use of a held-out calibration set, is lightweight and can be easily used in practice. The framework is further extended to allow distribution shift via the tracking and adaptation of the coverage level as time goes on, which can overcome the key limitations of the conformal methods that meet the exchangeability assumption. Risk-Controlling Prediction Sets (RCPS) extend the coverage guarantee to monotone risk functions (using arbitrary metrics like false negative rate, recall or user-specified task) objectives. RAPS (Regularized Adaptive Prediction Sets) is an enhanced efficiency of prediction set in classification by fining large set sizes, generating more informative uncertainty communication without compromising the coverage.

### 6.3 Uncertainty in Large Language Models

The uncertainty quantification frontier has reached a fresh and especially difficult stage owing to the explosive development of large language models (LLMs). LLMs are autoregressive models, meaning they make predictions at each token based on the final context per token, and thus the combining of token uncertainty with statement- or document-level confidence is not easily achievable. The average log-probability or perplexity of a generated sequence is an attempt by naive methods to use these values as an overall confidence measure, however they are ill-calibrated with respect to sampling

high temperature and they do not distinguish between the uncertainty that a given model has of the actual factual correctness of a claim and the uncertainty that a model has of the stylistic choice. Hallucination behavior - where LLMs write fluent statements with high confidence yet they are factually false - is a paradigmatic failure of self-awareness in epistemology with substantial implications on its use in information-critical systems.

A number of methodological strands are conceptualizing the uncertainty of LLM with greater complexity. The concept of semantic entropy proposed by Kuhn et al. (2023) sums the uncertainty on a semantic meaning level, by grouping several generations together based on their semantic equivalence and calculates entropy on meaning clusters as opposed to token sequences. This strategy is appropriate to recognize epistemic ambiguity regarding the factual material and not sensitive to the aleatoric fluctuation of expression. Verbalized uncertainty elicitation the most common spearheaded to motivate the LLM to self-report a belief in natural language has also been tested widely yet was found to correlate poorly with any empirical accuracy depending on the phrasing of the prompt and the scale of the model. To have coverage guarantees in question-answering tasks based on the model outputs, conformal prediction has been modified to use model outputs as nonconformity score generators. A potential opportunity to lessen epistemic uncertainty using the integration of retrieval-augmented generation (RAG) models alongside the UQ is the grounding of model results in external knowledge where uncertainty is used as a signal to activate retrieval augmentation.

## 7. Application and Deployment Planning as well as Reviewing Frameworks.

The process of applying UQ methods to research standards based on a real-world implementation contains a collection of pragmatic concerns that are usually overshadowed in the scholarly literature. The first of these is the computational cost factor of using approaches that entail more than one forward pass or ensemble inference. In latency-constrained systems, including autonomous driving perception, real-time medical monitoring, and interactive dialogue systems, it can be critically necessary to have a very small computational budget to estimate uncertainty, a constraint than single-pass methods (evidential DL, deterministic UQ, deep kernel learning) attempt to satisfy by producing the desired uncertainty estimates at low overheads. The compromise between computational efficiency and uncertainty quality is also application specific and has to be considered within the framework of the deployment restrictions of that particular issue.

Quantification of uncertainty in medical imaging has been used on various tasks such as detection of the disease, segmentation, and prognosis. A number of clinical trials indicated that the epistemic uncertainty maps have a significant relationship with the radiologist disagreement, lesion rarity and acquisition artifacts which imply that they

may have a role of decision support signal that attracts expert attention or increased imaging of such cases. The issue of high-quality uncertainty-annotated data, i.e. ground truth no longer being characterized by one label, but by the entire distribution of expert views, is being solved in multi-annotator studies and probabilistic ground truth models which explicitly model the heterogeneity of annotators. The regulatory implications are also serving to inform the ways in which uncertainty should be reported in cleared AI medical familiarity equipment with new recommendations on the FDA pointing at the necessity of disclosing information on the constraints of models and the degree of certainty.

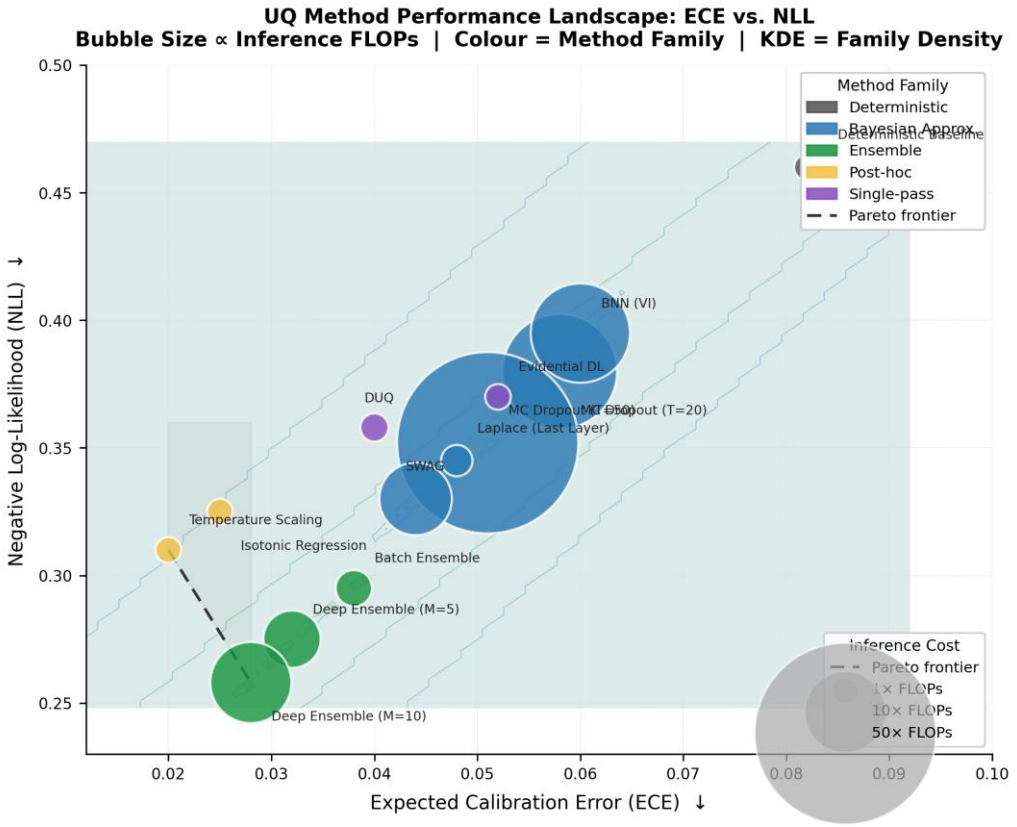


Fig.4 UQ Method Performance Landscape: ECE vs. NLL Bubble Chart with KDE Contours

Above Fig.4 is A pairwise bubble chart positioning 13 UQ methods in the two-dimensional evaluation space of ECE (calibration quality) vs. NLL (probabilistic fit), with bubble area proportional to inference FLOPs (computational cost) and colour encoding method family (Bayesian, Ensemble, Post-hoc, Single-pass). Family-level KDE contours reveal clustering structure. The Pareto frontier — methods that are non-dominated on both ECE and NLL simultaneously — is marked with a dashed curve.

Calibration and uncertainty emerged as the key evaluation criterion in the field of natural language processing as more knowledge-intensive tasks are conducted with the help of LLMs. Standardized test of factual accuracy and confidence calibration Benchmark suites such as Calibration Bench, TruthfulQA, and FAct Scoring offer domains, model scale and prompting strategy standard norms of factual accuracy and confidence calibration. Non-monotonic scaling as a result of calibration with model size has been discovered to be common and even is tolerated by large frontier models whose calibration may be worse than that of small model or even just memorization: the complex interplay between memorization, generalization and confidence learning during pretraining. Newer fine-tuning methods based on uncertainty-aware goal including reinforcement learning from human feedback (RLHF) variants that reward suitable hedging is beginning to be promising in improving the fine-tuning of instruction-tuned models without a negative impact on task performance.

UQ methodologies need special measures and standard of practices that are beyond generic measures of accuracy. Information about regression UQ is measured in a vast array of metrics, such as calibration, sharpness, and resolution decompositions (Uncertainty Toolbox, 2021). OpenOOD benchmark is an evaluation standard that standardizes out-of-distribution detection on various levels of OOD severity and semantic distances. WILDS offers distortion shift benchmarks obtained by real-life domain shift cases that allow to assess whether uncertainty techniques perform accurately by increasing uncertainty in conditions of realistic covariate shift. The use of the recently introduced HELM framework of LLMs and their uncertainty-aware extensions are towards a holistic evaluation that will focus on evaluating the performance, calibration, fairness and robustness all at once. The issue with UQ evaluation has always been that it is impossible to measure all desirable qualities of an uncertainty estimator with a single measure, which has led to the creation of multi-criterion evaluation methods that explicitly brings trade-offs between the natures of calibration, sharpness, computational cost, and OOD sensitivity to the surface.

## 8. Summary Tables

Table 1: Comparative Overview of Major Uncertainty Quantification Methods in Deep Learning

| Method / Framework | Uncertainty Type | Key Mechanism | Evaluation Metric | Primary Limitations |
|---|---|---|---|---|
| Monte Carlo Dropout (MC-Dropout) | Epistemic | Dropout at inference for approximate Bayesian posterior sampling | Predictive entropy, mutual information | Computationally expensive; requires tuning dropout rate |

| | | | | |
|---|---|---|---|---|
| Deep Ensembles | Epistemic + Aleatoric | Multiple independently trained networks; diversity via random init | NLL, ECE, Brier Score | High memory/compute; no shared representation |
| Heteroscedastic Neural Networks | Aleatoric | Predicts input-dependent output variance; NLL training objective | Calibration error, sharpness | Cannot capture model uncertainty; sensitive to outliers |
| Bayesian Neural Networks (BNNs) | Epistemic | Prior distributions over weights; variational inference or MCMC | ELBO, posterior predictive NLL | Intractable exact inference; high computational overhead |
| Normalizing Flows (UQ variant) | Aleatoric | Exact likelihood modeling via invertible transformations | Log-likelihood, CRPS | Limited to specific output distributions; complex training |
| Evidential Deep Learning | Epistemic + Aleatoric | Higher-order distributions (Dirichlet/Normal-Inverse-Gamma) over predictions | Uncertainty decomposition metrics, ECE | Training instability; requires careful loss design |
| Conformal Prediction | Aleatoric (coverage) | Distribution-free prediction intervals with marginal coverage guarantees | Coverage rate, interval width | Coverage but not probabilistic; exchangeability assumption |
| Stochastic Weight Averaging Gaussian (SWAG) | Epistemic | Gaussian approximation to posterior via SWA trajectory | ECE, NLL, accuracy under distribution shift | Gaussian assumption may be poor; memory intensive |
| Laplace Approximation (Post-hoc) | Epistemic | Second-order Taylor expansion around MAP estimate; Hessian computation | NLL, calibration on OOD | Full Hessian intractable; relies on local approximation |
| Bayesian Last Layer (BLL) | Epistemic | Bayesian treatment of final layer only; fixed feature extractor | ECE, AUROC on OOD detection | Ignores uncertainty in lower layers; limited expressiveness |
| Deep Kernel Learning + GP | Epistemic + Aleatoric | DNN feature extraction combined with GP kernel for posterior inference | GP marginal likelihood, posterior calibration | Scalability challenges; kernel design complexity |

Table 2: Application Domains, Benchmarks, and Emerging Frontiers in Deep Learning Uncertainty Quantification

| Application Domain | UQ Approach Used | Dataset / Benchmark | Key Finding | Open Research Challenge |
|---|---|---|---|---|
| Medical Image Segmentation | MC-Dropout + Ensembles | BraTS, LIDC-IDRI | High epistemic uncertainty correlated with expert disagreement in lesion boundaries | Aligning model uncertainty with clinical decision-making workflows |
| Autonomous Driving Perception | Heteroscedastic NNs, Evidential DL | KITTI, nuScenes, CARLA | Aleatoric uncertainty dominant in adverse weather; epistemic spikes near domain boundary | Real-time UQ under safety-critical latency constraints |
| Natural Language Processing / LLMs | Conformal Prediction, Token-level UQ | TriviaQA, MMLU, HaluEval | LLMs exhibit miscalibration; verbalized uncertainty poorly correlated with actual accuracy | Meaningful epistemic UQ for autoregressive generation without retraining |
| Drug Discovery / Molecular Properties | Deep Ensembles, Deep GP, BNNs | MoleculeNet, QM9, ChEMBL | Ensembles outperform single BNNs; uncertainty guides active learning loops effectively | Generalization of UQ to novel chemical scaffolds out-of-distribution |
| Remote Sensing / Earth Observation | Evidential DL, SWAG | UC Merced, BigEarthNet, SAR datasets | Epistemic uncertainty spikes at class boundary regions and underrepresented biomes | Handling spatiotemporal correlation in uncertainty estimates |
| Fault Detection in Industrial IoT | Bayesian LSTM, Conformal Prediction | CWRU Bearing, NASA CMAPSS | Predictive uncertainty increases prior to failure events; actionable early warning | Non-stationary data streams; concept drift and uncertainty recalibration |
| Reinforcement Learning / Robotics | Epistemic UQ for exploration, ensemble Q-functions | OpenAI Gym, DM Control, ROBEL | Uncertainty-driven exploration outperforms epsilon-greedy in sparse reward settings | Scalable epistemic UQ for high-dimensional continuous action spaces |
| Financial Risk Modeling | Heteroscedastic NNs, Deep Ensembles | S&P 500, Crypto markets, credit scoring | Aleatoric uncertainty captures volatility clustering; epistemic uncertainty flags tail-risk regimes | Regulatory interpretability requirements for UQ outputs in finance |
| Climate and Weather Forecasting | Probabilistic NNs, Normalizing Flows | WeatherBench, ERA5, CMIP6 | DL-based probabilistic forecasts approach NWP ensemble skill at lower compute cost | Calibrated extreme event uncertainty; physical consistency of probabilistic outputs |

| | | | | |
|---|---|---|---|---|
| Generative AI / Diffusion Models | Score-based UQ, Posterior sampling | CIFAR-10, ImageNet, CelebA-HQ | Diffusion models implicitly encode aleatoric uncertainty via stochastic sampling trajectories | Disentangling model uncertainty from data stochasticity in generative pipelines |
| Out-of-Distribution Detection | Energy-based, Mahalanobis distance, DUQ | OpenOOD benchmark, ImageNet-O, CIFAR-OOD | No single method dominates across all OOD severity levels; ensemble-based detectors most robust | Separating semantic OOD from covariate-shifted inputs; cost-effective deployment |

## 9. Open Research Frontiers

Although the UQ sphere has reached a certain degree of maturity, there are still several basic and practically valuable open issues that actively systematically research the entire community of deep learning. The first and possibly the main frontier is the uncertainty estimation that can occur in the context of distributional shift. Majority of the current UQ techniques are developed and tested based on the assumption that the input to the test is also sampled using similar distribution to that of the training data. Practically, though, there is a continuous drift in the distribution of inputs that are experienced at deployment as a result of domain shift, covariate shift and label shift. Techniques which generate uncertainty calibrated well to be used on training distribution can disastrously crash on shift, making overconfident predictions where pre-emption is most appropriate. The development of UQ tools with demonstrable or empirically sound properties under distribution shift is a not only open, but also actively tackled problem, which overlaps with the wider literature of domain generalization and domain adaptation.

The second frontier is regarding the communication between uncertainty reduction and scaling of the model. The behaviour of uncertainty estimates changes, not understood well, as the number of parameters and the size of the training set of deep learning models increase: using foundation models with hundreds of billions of parameters. Larger models usually come with a smaller irreducible test error, although their estimates of uncertainty in epistemics do not scale with their better accuracy, and their calibration characteristics are delicate to training and fine-tuning practices. Alterations in the temporal dynamics of epistemic uncertainty that are still to be described in detail have implications, as has been recently discovered, of grokking, in that a model keeps improving in generalization despite trains of thinking, after its loss has been obtained being finite. Moreover, the interplay between weight sparsity, quantization and uncertainty estimation in compressed models is a newer practical issue as deployment is considered via edge devices and resource constrained environments.

The third frontier is the criterion of uncertainty of generative models, such as diffusion models, variational autoencoders, and large language models. The resulting architectures are expressing the aleatoric variability in the output space using a diffusion model that samples an image, but a model where all of the outputs of the model have the same image no matter the latent code might be showing collapse and not certainty. Coming up with conceptually rigorous systems of quantifying epistemic uncertainty in neural networks, such as the distinction between the uncertainty of the model more generally with the data manifold or natural variation in the target distribution, is a highly complex issue with critical consequences in drug design, scientific simulation and creative AI.

Lastly, the correspondence between the human intuitions concerning the uncertainty and the quantities that are approximated by the computational UQ approaches are not perfect and under theorized. Human uncertainty is situational, expressive and action related in manners mathematical entropy or predictive variance may not entirely describe. The interdisciplinary challenge, that involves cognitive science, decision theory, and human-computer interaction among machine learning, is associated with designing interfaces that convey information about model uncertainty to human operators in a manner that can be understood (and acted on), but will not be misinterpreted. Explainable AI (XAI) research has also started to overlap with UQ studies fruitfully, where the goal of providing uncertainty explanations is based on confidence levels derived on particular input features or training examples as opposed to an abstract distributional summary. Such anthropocentric view of uncertainty communication, in turn, is bound to gain more significance with the implementation of UQ mechanisms in consequential sociotechnical systems.

## 10.Conclusion

Uncertainty quantification in deep learning is at the cross-roads of the theory of statistics, the methodological requirements of computing, and the practical demands of the deployment of reliable AI. The underlying difference between aleatoric and epistemic uncertainty offers a fruitful conceptual framework through which it can be said to be what a model predicts and to what extent it is of value that its predictive representations can be trusted and to what circumstances the trust is appropriate. This chapter has navigated the terrain of the UQ since its probabilistic roots up to the participation of the key methodological traditions, which are Bayesian neural networks, deep ensembles, heteroscedastic models, MC Dropout, evidential deep learning and conformal prediction, to the current questions at the boundaries of the large language model, generative AI and the distribution shift.

The discipline has achieved splendid endeavors in coming up with computationally affordable approaches that deliver viable uncertainty estimacy in a broad spectrum of

setups and practice-areas. Deep ensembles are an effective practical motivation; conformal prediction has hard coverage guarantees; and the Bayesian model has been stimulating novel approximations that trade between theory and computability. Simultaneously, the difference between theoretical principles and practical reality is still enormous: the approaches, which work well on standardized models, may silently undergo the distribution shift, and even the calibration of the state-of-the-art models can be not perfect even with the friendly circumstances.

The reason why the way to a genuine collection of trustworthy deep learning needs not only improved uncertainty estimators, but also improved uncertainty assessment, communication, and action structures in high-stakes situations. The conceptual quantification of epistemic and aleatoric uncertainty will not become an optional addition, but a core requirement to be able to make responsible use of artificial intelligence as it finds its increasing uses in problems where mistakes can be very costly. The level of intellectual momentum apparent in the existing literature, such as Bayesian deep learning, frequentist coverage theory, information theory and behavioural interaction, is indicative that the discipline can address this challenge, as long as the researchers and practitioners adopt stringent criteria concerning evaluation, reproducibility, as well as honest disclosure of the shortcomings of current methods.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 4: Probabilistic Deep Learning: Bayesian Neural Networks and Deep Ensembles

## Abstract

The use of deep learning systems in high-stakes settings, such as clinical decision support, autonomous navigation, financial risk modelling, and safety-critical infrastructure management, has made the need to have models that, in addition to high predictive accuracy, generate well-calibrated, interpretable and reliable uncertainty estimates. Probabilistic deep learning answers this call by offering principled models in which neural networks can be described to represent epistemic uncertainty due to scanty or unclear training data, and aleatoric uncertainty due to stochastic mechanisms that gave rise to observations. This chapter will give an in-depth academic analysis of the two leading paradigms in probabilistic deep learning the example of Bayesian Neural Networks (BNNs) and Deep Ensembles. We use a first-principles derivation of the theoretical foundations of each approach, provide a survey of the landscape of approximate inference methods developed to make Bayesian inference scalable and manageable, analyze the new literature on the interrelations and differences between these two paradigms, and review new developments such as scalable Bayesian inference methodology, functional uncertainty quantification and integrating probabilistic methods and large-scale learnt representations. At the end of the chapter, the paper will end by giving a synthesis of open challenges and a future outlook regarding the the path of the trustworthy probabilistic deep learning.

## 1. Introduction

The current state of modern deep learning has demonstrated an impressive empirical performance in an extremely diverse set of perceptual and sequential prediction problems but the overconfidence of conventional deterministic neural networks is a fail mode which is both persistent and consequential. Even when maximum likelihood estimation has been used to train a deep neural network to make softmax predictions

with high confidence, the resulting confidence score is not a principled prediction of the epistemic status of the model--instead, it, at best, is a heuristic score determined by the architecture and training process, and at worst, it is a dangerously false indicator of accuracy. The awareness of this is propelling a revitalization of probabilistic modelling in the deep learning community, with roots in the classical notions of Bayesian statistics, information theory, statistical physics, and approximate inference that have long been the focus of thinking but were believed at scales that are computationally infeasible in the current neural networks.

Bayesian learning neural networks The intellectual history of Bayesian learning of neural networks dates back to the original work of MacKay (1992) and Neal (1995) who realised that when prior distributions are imposed on the weights of a neural network and the posterior distribution of the weight is computed with the help of the observed data, prediction would be made in a way that marginalises the probability distribution of the model uncertainty, not to the point estimate. This philosophically consistent and theoretically valid Bayesian treatment was mostly replaced by the 1990s and 2000s as neural networks deepened and became wider, and the problem of the impossibility of exact posterior computation rendered the workflow to be unfeasible. Since 2015, the renewed attention has been fuelled by both the convergence of scalable variational inferences algorithms and the development of approximate inferences algorithms that take advantage of modern auto differentiation models and by the practical need to measure uncertainty in deployed machine learning systems.

Otherwise, similarly to the Bayesian tradition, the ensemble approach to uncertainty estimation has proved to be an academically and practically interesting alternative. The original observation, which is rigorously defined by Lakshminarayanan, Pritzel, and Blundell (2017), is that when trained on different random initialisations and the average of their prediction is calculated, the resulting uncertainty estimates of the approach in question always perform better than single-model approaches such as approximation Bayesian estimation in multiple aspects: calibration quality, out-of-distribution detection, and downstream task performance. Deep ensembles have become so easy and useful that they have become de facto whose benchmark all other methods of uncertainty estimation are gauged.

The chapter is structured in such a way that it gives a systematic treatment to the two paradigms. Section 2 elaborates the theoretical background of Bayesian neural networks, which reviews the prior-posterior viewpoint, predictive distribution, and the breakup of uncertainty into epistemic and aleatoric parts. Section 3- reviews the pastures of approximation inference techniques that have rendered BNNs amenable such as variational inference variants, Laplace approximations, Markov chain Monte Carlo methods, and stochastic weight averaging methods. Section 4 looks at deep ensembles, and how they are built, the theoretical definitions of their uncertainty properties and their

algorithmic advances have decreased their computational burden and retained their uncertainty quality. Section 5 summarises the literature regarding relationships between Bayesian and ensemble views, discussing functional diversity, functional use of loss surface geometry and the recent theoretical synergies. In section 6, the authors discuss the combination of probabilistic models with pre-trained foundation model and learning a big representation. Section 7 points out the challenges which are open and sets the course of further research. There are two summary tables, which are structured compare and contrast approaches to important methods.

## 2. Bayesian Neural Networks basics theory.

### 2.1 Bayesian Framework of the Neural Networks.

Bayesian interpretation of neural networks as specifying a prior distribution p (w) along the weights w of the network, that is, capturing information about values of the parameters in advance of observing data. On a set of input-output pairs, D = {(x_i, y_i)}_{i=1}^N Nuptess the Bayesian beliefs are updated by Bayes' theorem to a posterior probability p(w|D) given w and the likelihood p(D|w). Not the posterior but the posterior predictive distribution is the object of interest here, so that with a test input x* it would appear as $p(y^*|x^*, D) = \int p(y^*|x^*, w)\, p(w|D)\, dw$.. This intrinsic marginalises forecast through the full posterior distribution, which is an effect of weighing each instance of a model by its likelihood, based on observed information. It is a predictive distribution, which concentrated around a particular weight configuration is usually more diffuse like the prediction of any given weight configuration and embodies the entire epistemic uncertainty due to incomplete information.

The difficulty of the practical problem that has stimulated decades of research is the intractability of both the posterior p (w|D) and the marginalisation integral of the predictive distribution. In any but the most basic models and data, the posterior on neural network weights does not have a closed form representation since the directed relationship between weights and predictions (which is a result of the neural network function) is nonlinear and non-conjugate. The evidence of the model $p(D) = \int p(D|w)\, p(w)\, dw$, known as the normalising constant, is also uncomputable, and it does not allow one to directly compute the posterior. Therefore, every practical Bayesian neural network technique is based on approximation, whether approximating the posterior using some computeable surrogate distribution q(w) (variational inference), using samples drawn under some distribution that tends to convergence to the posterior (Markov chain Monte Carlo), or approximating the posterior about the mode of the posterior (Laplace approximation). The quality of uncertainty estimates and

computational practicality of the approach is basically dependent on the nature of such approximations.

### 2.2 Epistemic and Aleatoric Uncertainty

The principled decomposition of total predictive uncertainty into its epistemic and aleatoric parts is also conceptually significant to trustful deep learning of the Bayesian framework. Epistemic uncertainty or model uncertainty or reducible uncertainty is brought about by the fact that the model may have an incomplete knowledge of the actual data-generating process and can in principle be decreased by obtaining more informative observations. Within the Bayesian model, epistemic uncertainty is represented by the dispersion of the posterior distribution over the model parameters: areas of the input space which are distant to the training data, or parts of the input space which have divergent predictions when the parameters are varied will have high epistemic uncertainty. Aleatoric uncertainty, in its turn, is a manifestation of irreducible randomness that is present in the process of observation, noise in sensor measurements, ambiguity in labelling conventions, or actual stochasticity in the phenomenon being predicted, and cannot be mitigated no matter how large the dataset is. This decomposition was formalised by Kendall and Gal (2017) to deep learning applications and demonstrated that the predictive variance can be broken down into the sum of expected data noise (aleatoric) and variance of the expected prediction across the posterior (epistemic), as well as that each component informs practitioners who deploy models in real applications about something meaningful to do.

Such a difference in the types of uncertainty has far-reaching practical consequences. Aleatoric uncertainty of imaging, which can be created by artefacts of medical imaging, or a truly ambigious pathology, can be incapable of being reduced by refining models, and this should be communicated to clinicians so that he or she can make the correct decision. Epistemic uncertainty, in its turn, is informative to active learning, i.e. to which unlabelled examples to minimise model uncertainty would it be labelled; to out-of-distribution detection, where high epistemic uncertainty on unseen ex samples would indicate that the model is being run in a region it does not predict very well. This separation of quantifying and communicating both forms of uncertainty is a unique benefit of the probabilistic deep learning methods properly run as opposed to the deterministic methods, and one of the main reasons why the research community has invested so heavily in making them computationally amenable.

### 2.3 Previous Specification and inductive Biases.

Prior distribution p(w) specification is a design choice in Bayesian neural networks which has been given renewed theoretical focus and practical emphasis in recent years. The most widely used and simplistic prior is the isotropic Gaussian p(w) = N(0, σ 2 I) which would in the limit of maximum a posteriori (MAP) estimation amounts to the L2 regularisation, and which can be simply described as one preferring short sequences of a small magnitude on a parameter. Nonetheless, isotropic Gaussian priors have been noted to have the following shortcomings: these priors apply the same strength of regularisation to all layers and types of parameters, are incompatible with the highly structured sparsity that typically appears in well-trained neural networks, and scale badly to the underlying geometry of the actual posteriors. More expressive prior choices are hierarchical or empirically-motivated Bayes prior, which can predetermine hyperparameters through maximising marginal likelihood, horseshoe priors which support sparsity and large weights, which capture important features, and structured priors, which encode architecturally-motivated inductive biases, like translation equivariance or hierarchical feature composition.

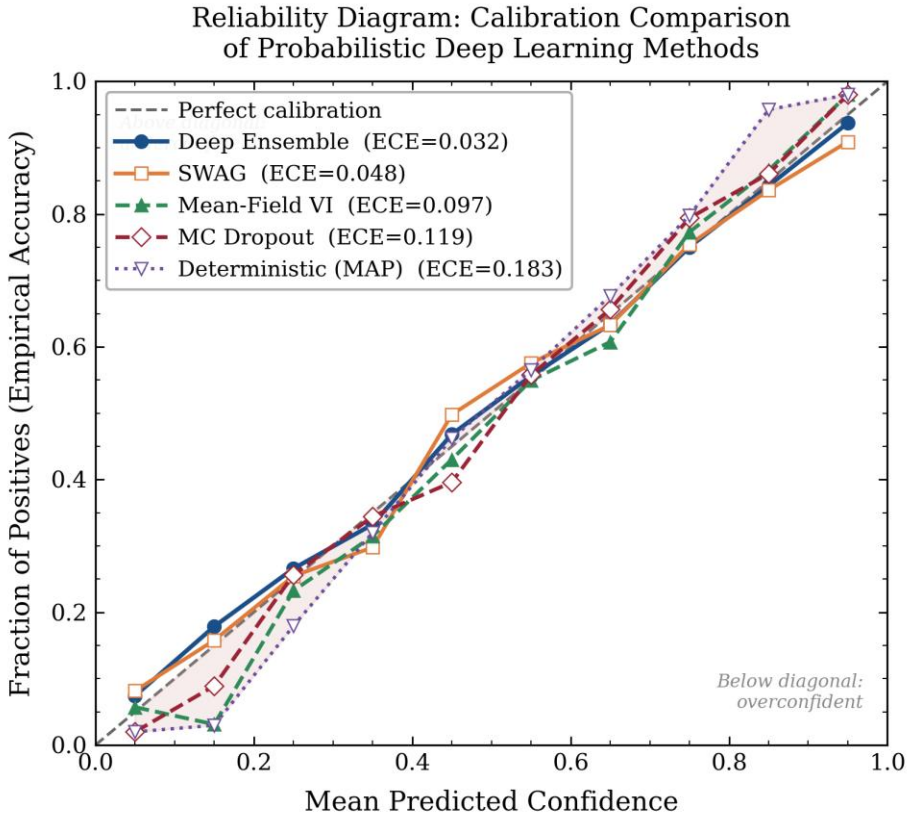


Fig.1 Reliability Diagram (Calibration Curves)

Fig.1 shows empirical accuracy against mean predicted confidence across 10 equally spaced bins for five methods. A perfectly calibrated model traces the diagonal. Expected Calibration Error (ECE) is reported for each method in the legend. Deep Ensembles (ECE = 0.032) and SWAG (ECE = 0.048) lie near-perfect calibration, while the deterministic MAP baseline (ECE = 0.183) falls well below the diagonal, revealing systematic overconfidence. The pink shaded region highlights the overconfidence gap of the MAP model.

Another very dynamic field of research has been the specification of priors in functional space as opposed to weight space, or to put it another way, the so-called functional priors or priors on the input-output maps caused by the network, which is also known as priors over functions. The incentive is that weight-space priors which have ostensibly reasonable formulations can cause the pathological functional priors of deep networks: such as the standard weight-space priors of deep networks drive functions to be concentrated on the boundaries of the output simplex, predicting with maximum certainty, which is counterintuitive in the intuition that the prior should be able to represent true ignorance before data. The methods of specifying functional priors directly, either by Gaussian process correspondences or by approximate inference in the functional space, have been developed in work by Flam-Shepherd, Requeima, and Duvenaud (2017), Hafner et al. (2019), and later contributions and have been shown to have better uncertainty calibration as a side-effect. This research is related to Botanian neural networks as a generalization of the infinite-width neural tangent kernel and Gaussian process literature, which offers theoretical support to the connection between prior specification, architecture and the quality of uncertainty estimates.

## 3. Bayesian Neural Network Methods of Approximate Inference.

### 3.1 Variational Inference

Variational inference transforms the intractable posterior inference problem into an optimisation problem by introducing a parameterised family of approximate distributions $q_\varphi(w)$ and seeking the member of this family that minimises the Kullback-Leibler divergence $KL[q_\varphi(w) \,||\, p(w|D)]$ from the true posterior. This minimisation is equivalent to maximising the Evidence Lower BOund (ELBO): $L(\varphi) = E_{q_\varphi(w)}[\log p(D|w)] - KL[q_\varphi(w) \,||\, p(w)]$, which balances data fit (the likelihood term) against prior regularisation (the KL divergence term). The seminal Bayes by Backprop algorithm of Blundell et al. (2015) demonstrated that this objective could be efficiently optimised using the reparameterisation trick—expressing the random weights as deterministic transformations of noise samples $w = \mu + \sigma \odot \varepsilon$, $\varepsilon \sim N(0,I)$—enabling unbiased gradient

estimates with respect to the variational parameters (μ, σ) and integration of variational BNN training within standard stochastic gradient descent frameworks.

The most popular variant of the variational inference family is the mean-field approximation, which makes the assumption that the posterior $q_\varphi(w) = \prod_i q(w_i)$ is fully factored; in other words, the individual distributions on each weight are independent. Although computationally easy the number of variational parameters depends only linearly on the number of network parameters, the mean-field assumption has been shown to underestimate posterior uncertainty in a systematic way when it ignores inter-parameter correlations which can be very large in well-trained deep networks. Full-covariance variational inference represents these correlations at the cost of a quadratic scaling in the number of parameters, which is prohibitive with modern architecture. Structured approximations, which lie intermediate between exact and heuristic approaches, have been invented such as Kronecker-factored (KFAC) covariance approximations, rank-one perturbation methods, inductive-point representations, adopted over sparse Gaussian process literature; these methods partially approximations of the correlations at affordable computational cost. The Flipout estimator of Wen et al. (2018) estimates gradient variations between examples with a mini-batch more successfully to control the high variance of gradient estimates, by decorrelating the perturbations placed on different examples using pseudo-independent random sign matrices, with no change to the underlying variational objective.

Other than the mean-field family, normalising flows have been used to make expressive approximate posteriors via successive transformation of a simple base distribution by a series of invertible, differentiable transformations. Flow-based posteriors have the advantage of being able to represent arbitrary distributions with multimodal and heavy-tailed shapes, being able to overcome the elementary limitations of Gaussian means-field approximations to expressivity. This method was demonstrated by Louizos and Welling (2017) who built flexible approximation posteriors based on Householder transformations, and the estimate of marginal likelihood was better compared with baselines based on mean fields. The practical implications such as the computational costs of flow-based posteriors in terms of the number of parameters and inferencing time, and the creation of normalising flows that support high-dimensional weight space are currently being studied. Recent methods on the Riemannian normalising flows which preserve the natural geometry of the parameter space provide a promising direction where more expressiveness is achievable at a cost less proportional to the expressiveness.

### 3.2 Markov Chain Monte Carlo Methods

Markov chain Monte Carlo (MCMC) algorithms give asymptotically precise posterior samples, built by means of a Markov chain with a simplex which is the target posterior

p(w|D). The classical MCMC methods such as Metropolis-Hastings algorithm and Gibbs sampling are computationally infeasible on modern neural networks with millions of parameters because they either need to evaluate the entire dataset or calculate a proposal distribution which is not scalable to network size. Resolving this issue, Stochastic gradient Langevin dynamics (SGLD) presents by Welling and Teh (2011) injects suitably scaled Gaussian noise into stochastic gradient descent updates, and thus the optimisation algorithm is converted into a sampling algorithm, the samples of which will converge to the posterior in the limit. It has been found that the relationship between SGLD and approximate Bayesian inference can be highly fruitful, and that broad variants of stochastic gradient MCMC methods such as SGHMC, SGNHT, and preconditioned versions can incorporate second-order curvature information in order to mix easier.

Irrespective of these developments, full MCMC posterior inference of current deep networks is also computationally expensive. The high-dimensional complex loss landscapes of deep networks have extremely long mixing times of Markov chains, and it may take long before samples are approximately drawn out of the posterior. Cyclical stochastic gradient MCMC is a method that overcomes this drawback and makes use of cyclic learning rate schedules that alternate between exploration phases, during which the chain mixes and explores the posterior, and exploitation phases, during which high-likelihood regions are traversed. Empirically verified by Zhang et al. (2020), this method allows computing various approximate posterior samples with realistic training budgets and has been found to generate adversarial uncertainness estimates on conventional benchmarks. The correlation between MCMC samples gathered in SGD trajectories and the actual posterior samples is an active field of theoretical research, and it is related to the loss surface topography, sharpness of the local minima, and the generalisation of stochastic gradient algorithms.

### 3.3 Laplace Approximation and Its Modern Variants

One of the oldest approximations of Bayesian inference, which can be analysed analytically, is the Laplace approximation, which approximates the posterior as a Gaussian with mean set to the maximum a posteriori (MAP) estimate with covariance equal to the inverse Hessian of the negative log-posterior at the maxima a posteriori. Although mathematically simple with small model sizes, and often yields a more intuitive estimate in high-dimensional models, the Laplace approximation has two major issues in deep learning, where computationally infeasible answers are required: the MAP estimate does not necessarily identify an actual mode of the posterior distribution, especially on flat loss landscapes and low-curvature loss landscapes that contain many parameter configurations with the same likelihood, and calculation or storage of the full Hessian is infeasible with networks of millions of parameters. The Kronecker-factored

modelling of the Hessian of Ritter, Botev, and Barber (2018) overcomes the second challenge of expressing the Hessian as a Kronecker product of smaller matrices, with each layer and the correlation between layers not considered, to make the Hessian cheap to invert and store.

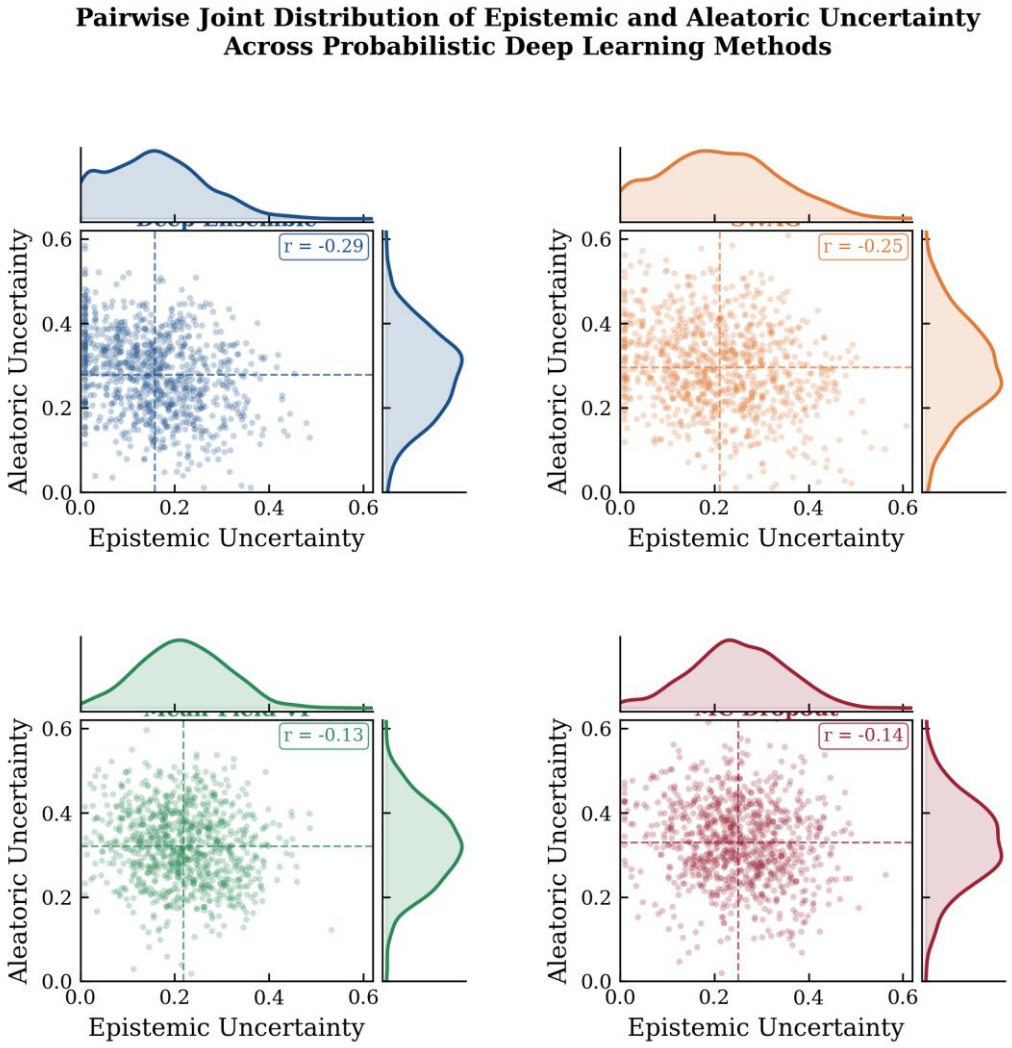


Fig.2 Pairwise Scatter: Epistemic vs Aleatoric Uncertainty

Fig.2 shows A 2 × 2 panel figure showing the joint distribution of epistemic and aleatoric uncertainty for 900 test predictions per method. Marginal KDE curves decorate each axis. The Pearson *r* annotated in each panel reveals that Deep Ensembles and SWAG exhibit negative correlation ($r \approx -0.30$ to $-0.25$), meaning these methods correctly assign

lower aleatoric confidence when epistemic uncertainty is high — a key indicator of a well-calibrated probabilistic model.

The recent influential contribution by Daxberger et al. (2021) has contributed hugely to bringing back interest in the Laplace approximation, with the Laplace Redux framework showing that out-of-proximity application of the Laplace approximation to the final layer of a pre-trained neural network (the penultimate layer effectively a fixed feature extractor) with only the classification head subject to Bayesian inference can offer uncertainty estimates similarly competitive to methods that are much more computationally expensive to run). Such last-layer Laplace approximation is especially appealing in situations with large pre-trained models, in which retraining with full Bayesian inference would be prohibitively costly. This insight has been extended in the subsequent work to a family of subset Laplace approximations using Gaussian posterior applied to a chosen subset of parameters that are most likely to be of interest in the uncertainty context, giving a range of trade-off between the quality of error and the cost of computation.

### 3.4 Probabilistic and Stochastic Weight Averaging.

The training process, proposed by Izmailov et al. (2018), called Stochastic Weight Averaging (SWA) is a training procedure that averages neural network weights over time, observed along the path of SGD, showing as again that simple averaging process consistently results in wider, flatter minima than standard SGD and produces better generalisation. Maddox et al. (2019) have presented the probabilistic extension, the Stochastic Weight Averaging-Gaussian (SWAG) estimating a Gaussian distribution on the SGD trajectory by capturing both the first and second moments of the parameter trajectory and building a low-rank plus diagonal covariance model. When this Gaussian is sampled at test time and predictions averaged over samples give uncertainty estimates which have been shown to be competitive with deep ensembles on a variety of standard benchmarks, an impressive conclusion given that computation requirements are much lower than the expense of training a collection of independently trained models. SWAG is a successful restatement of the stochastic exploration of the loss surface when SGD is used as approximate posterior sampling, which is a successful attempt to bridge the optimisation-focused and inference-focused interpretations of training neural networks.

## 4. Deep Ensembles Construction, Theory and Algorithms Advances.

### 4.1 Foundations of Deep Ensembles

Lakshminarayanan, Pritzel, and Blundell (2017) systematise the deep ensemble methodology, in which a set of M neural networks are initialised with various random seeds and adversarial training on gradient sign method perturbations in order to improve the diversity and expressiveness of these neural networks. The predictive distribution has been constructed by taking the average of the softmax outputs of individual ensemble members to produce a mixture of softmax distributions that contains not only the average confidence of individual ensemble members but also the disagreement between them. Operationalisations of the uncertainty of an ensemble include predictive entropy, mutual information between model parameters and model predictions, or the variance in predicted probabilities amongst ensemble members all of which represent slightly different facets of the collective uncertainty of an ensemble.

Both empirical and theoretical aspects The empirical effectiveness of deep ensembles to yield good calibrations of uncertainty estimates and reliably identify out-of-distribution inputs was intuitive when it was originally announced, and somewhat theoretically puzzling. In contrast to Bayesian neural network, deep ensembles do not have a literal interpretation as an approximation to the Bayesian posterior-4, the training algorithm does not require any explicit specification of the prior, no marginalisation objective, no variational approximation of the evidence lower bound. However, empirically, groups are always better at performing than theoretical-based techniques. The other theory that has been put forward by Wilson and Izmailov (2020), made theoretical rationalisation quite compelling, by showing that deep ensembles can be thought of as doing Bayesian model averaging across a collection of hypothesis classes, each basin of attraction in the loss landscape, with each random initialisation resulting in a different basin. In this perspective distributions of ensembles do not approximate a prior on one model class but instead they marginalise in a more general space of hypotheses linearising with multi-modal Bayesian inference which single-basin methods can not cover.

### 4.2 Efficient and Scalable Ensemble Variants.

The main weakness of a typical deep ensemble is its computational inability: it is necessary to run M different models to obtain a list of M parameter estimates used for prediction at test time, and to calculate the prediction requires M forward passes of each prediction. This overhead has encouraged a deep tradition of studies on ensemble approximations which maintain the quality of uncertainty of conventional ensembles with smaller computed cost. Wen et al. (2020) suggest Batch Ensemble restating each member of the ensemble as a rank-one perturbation - a pair of input and output scale vectors - of the shared weights. The effective weight of the member m is defined as W m = W o (r m s m ) and member specific vectors, r m and s m are used and the entire ensemble can be calculated in one forward run by connecting the tiles together and

applying member specific perturbations. This method is able to perform almost standard-ensemble performance at a fraction of the cost in memory and computational time.

Packed Ensembles Packed Ensembles, suggested by Laurent et al. (2023) Applies to the network via grouped convolutions that divide the feature representations of the network into M groups with each group representing an ensemble member. The point is that grouped operations of convolution, which are highly optimised on a modern gpu, can execute the forward pass of all the ensemble members at once at no more cost than a single model and maintain the diversity properties of standard ensembles. Empirical analysis of the standard image classification and regression benchmarks have shown that Packed Ensembles can be as accurate and as well calibrated as a standard deep ensemble at 1/M the cost of computation and memory, and hence ensemble-quality uncertainty estimation is now available in resource-constrained deployment environments.

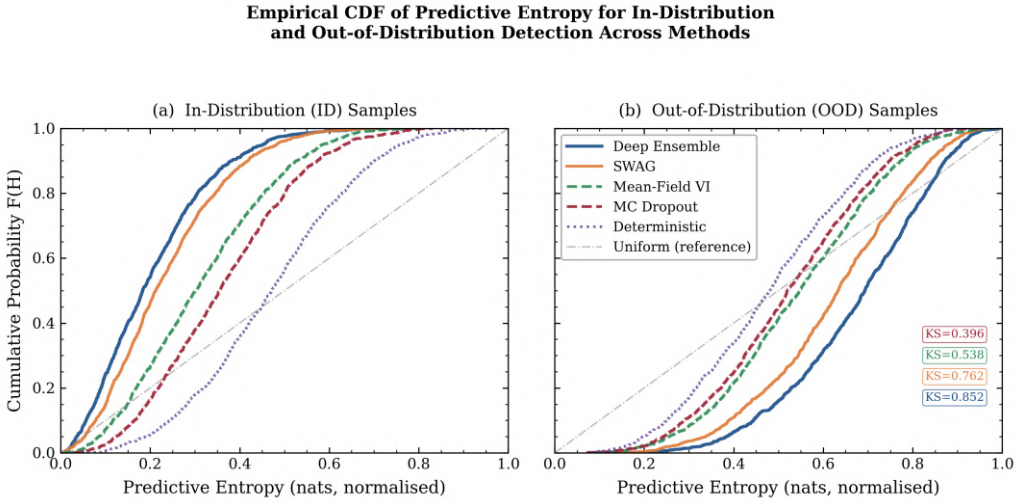


Fig.3 Empirical CDF of Predictive Entropy (ID vs OOD)

Above fig.3 is a Two-panel ECDF plot comparing normalised predictive entropy distributions on in-distribution (panel a) and out-of-distribution (panel b) test sets. A steep ID curve (low entropy) coupled with a flat OOD curve (high entropy) signals strong OOD-detection capability. Kolmogorov–Smirnov (KS) statistics between each method's ID and OOD distributions are annotated on panel (b); Deep Ensembles achieve the largest KS statistic, confirming superior distributional separation.

Another set of methods aiming to distil the uncertainty information of an ensemble is functional ensemble distillation and amortised ensemble methods. Ensemble distillation makes a student network predictive to an ensemble of networks, and would usually optimise the KL divergence between the distribution of predictions given by the student

and the predictive mixture distribution of the ensemble. The resulting student model is able to create predictions, which concur with the ensemble uncertainty at the cost of running a single forward pass, and, therefore, can be deployed efficiently with significant cost savings to the calibration advantage of running the full ensemble inference. This motivates Malinin et al. (2020), who utilized Prior Networks because they act as the student architecture, and attained competitive OOD detection and calibration indicators with significantly lower test-time overheads.

### 4.3 Diversity, Disagreement and the Role of Loss Landscape Geometry.

The conceptual insight into the mechanism of the effectiveness of ensembles, as well as methods of effectively building them, has been enriched significantly during the last years because of the research regarding the connection between diversity, the geometry of loss landscapes, and the quality of uncertainty. One of these observations is that the diversity of ensemble members, which can be defined in terms of prediction disagreement on the training set or distance between them in weight or function space, is a central variable in ensemble quality: ensembles where members lie near the same solution to the problem are prone to having correlated errors and contribute little to the overall calibration and OOD detection, but those where members lie in well-separated basins of attraction always achieve better calibration and OOD detection. The initialisation strategy, stochasticity of training process, learning rate and explicit mechanisms used by diversifying nature are determinants of the degree of diversity.

The empirical research of Fort et al. (2019) on geometry of neural network loss surfaces that are explored by various members of the ensemble revealed that even standard deep ensemble members, despite their equal architecture and training, always reach different basins of the loss surface that are divided by a high-loss barrier. It is this multi-modal geometry of the surface of losses, instead of a unimodal geometry about one global minimum, that forms the geometric underlay of ensembles that characterize the posterior diversity. In comparison, several trajectories of MCMC chains initialised in close proximity to an initial position are likely to be at the same basin generating locally diverse but globally similar samples, which in part explains the performance of deep ensembles over MCMC-based approaches despite the latter having better theoretical underpinnings. These lessons have inspired new training approaches which directly motivate list members to search different basins, including Diverse Ensemble Training approaches which use pairwise losses on disagreements and particle-based optimisation like Stein Variational Gradient Descent used on ensemble of neural networks.

## 5. Summary Tables

Table 1. Comparative Overview of Bayesian Neural Network Inference Methods

| Inference Method | Theoretical Basis | Computational Cost | Scalability | Uncertainty Quality |
|---|---|---|---|---|
| Markov Chain Monte Carlo (MCMC) | Exact posterior sampling via ergodic Markov chains; gold standard for posterior fidelity | Extremely high; $O(N^2)$ per step in naive implementations | Scales poorly beyond small networks; prohibitive for modern architectures | Excellent; asymptotically exact posterior representation |
| Hamiltonian Monte Carlo (HMC) | Exploits gradient information to propose efficient transitions; reduces random walk behavior | Very high; requires full gradient computation and leapfrog integration | Limited scalability; impractical for networks with >10M parameters | Very high fidelity; superior mixing properties over standard MCMC |
| Mean-Field Variational Inference (MFVI) | Assumes diagonal posterior covariance; minimizes KL divergence from factored Gaussian | Low to moderate; comparable to two standard forward passes | Highly scalable; applicable to large networks and datasets | Underestimates posterior variance; overconfident on OOD data |
| Full-Covariance Variational Inference | Captures inter-parameter correlations via full-rank covariance approximation | High; covariance matrix scales quadratically with parameter count | Moderate; feasible for medium-sized networks with structured approximations | Better calibrated than MFVI; captures parameter correlations |
| Stochastic Weight Averaging Gaussian (SWAG) | Fits Gaussian to SGD trajectory statistics; low-rank plus diagonal covariance model | Low; minimal overhead over standard SGD training | Highly scalable; demonstrated on ResNets, Transformers | Competitive with deep ensembles; well-calibrated uncertainty estimates |
| Laplace Approximation | Second-order Taylor expansion around MAP estimate; Gaussian centered at mode | Moderate; requires Hessian or Fisher information matrix approximation | Scalable with Kronecker-factored (KFAC) approximations | Reasonably calibrated; performance depends on curvature quality |
| MC Dropout (Variational) | Dropout as approximate variational inference; closed-form KL divergence term | Low; standard dropout without additional parameters | Excellent; applicable to any architecture with dropout layers | Moderate; tends toward overconfidence; sensitive to dropout rate |
| Flipout Perturbation | Decorrelated weight perturbations using pseudo-independent random signs | Moderate; 2x forward pass overhead per sample | Good; demonstrated on convolutional and recurrent architectures | Improved over standard VI; reduced variance in gradient estimates |
| Natural Gradient VI (VOGN) | Exploits natural gradient geometry for efficient posterior updates | Moderate; requires online Fisher | Demonstrated on CIFAR and ImageNet- | Well-calibrated; superior convergence |

| | | | | |
|---|---|---|---|---|
| | | information estimation | scale experiments | properties over standard VI |
| Normalizing Flow Posterior | Transforms simple base distribution through invertible neural mappings | High; flow complexity adds substantial computational overhead | Limited; flow architectures constrain posterior expressiveness in high dimensions | High expressiveness; can capture multimodal and complex posteriors |
| Expectation Propagation (EP) | Message-passing algorithm targeting moment-matching approximation | Very high; iterative message passing across likelihood and prior factors | Limited scalability; convergence not guaranteed in complex models | High-quality marginal approximations when convergence is achieved |
| Last-Layer Bayesian Approximation | Applies Bayesian treatment only to final classification or regression layer | Very low; Bayesian inference limited to small parameter subspace | Excellent; applicable to any pre-trained backbone network | Limited; captures only output-layer uncertainty, not representational uncertainty |

Table 2. Comparative Overview of Deep Ensemble Methods and Variants

| **Ensemble Variant** | **Diversity Mechanism** | **Training Protocol** | **Inference Strategy** | **Performance Characteristics** |
|---|---|---|---|---|
| Standard Deep Ensembles | Random initialization with different seeds; stochastic gradient noise provides implicit diversity | Independent training of M identical architectures with fresh random seeds | Predictive averaging of M softmax outputs; variance across members estimates uncertainty | State-of-the-art calibration; superior OOD detection; 3–5x computational cost over single model |
| Batch Ensemble | Rank-one perturbation matrices applied per ensemble member; shared base weights | Single forward pass with member-specific multiplicative perturbations; memory efficient | Simultaneous inference across members in single batch; fast test-time computation | Near-ensemble performance at fraction of cost; slight accuracy trade-off versus standard ensembles |
| Hypermodel Ensembles | Amortized hypernetwork generates diverse parameter distributions conditioned on index | Joint training of hypernetwork and primary network; single training procedure | Sample model parameters from hypernetwork; stochastic at test time | Competitive uncertainty quality; significantly reduced parameter storage overhead |
| Monte Carlo Dropout (Ensemble View) | Stochastic dropout masks induce different effective sub-architectures per forward pass | Standard dropout training; no special ensemble training protocol required | T stochastic forward passes with dropout active; predictive mean and variance aggregated | Convenient and scalable; underperforms calibrated ensembles; sensitive to |

| | | | | |
|---|---|---|---|---|
| | | | | dropout configuration |
| Snapshot Ensembles | Cyclic learning rate schedule drives parameters to multiple local minima sequentially | Single training run with cosine annealing restarts; snapshots taken at cycle ends | Average predictions from checkpoints; no additional storage of separate models | Training-efficient; diversity limited by proximity of local minima explored during optimization |
| Loss-of-Plasticity Ensembles | Periodic network reinitialization maintains diverse parameter exploration across training | Iterative training with selective weight perturbation; prevents premature convergence | Aggregate predictions from reinitialised subnetworks at evaluation time | Maintains diversity over long training horizons; addresses catastrophic forgetting in online settings |
| Functional Ensemble Distillation | Knowledge from ensemble members distilled into single probabilistic student network | Distillation loss matches student predictive distribution to ensemble distribution | Single forward pass through distilled model; uncertainty from student's output distribution | Inference efficiency of single model with approximated ensemble uncertainty quality |
| Packed Ensembles | Group convolutions implement M ensemble members within single network architecture | Train single model with grouped convolutions; each group represents distinct member | Single forward pass exploiting parallelism of group convolution operations | Near-standard-ensemble performance; inference cost comparable to single model; GPU-efficient |
| Masksembles | Binary masks applied to shared parameters; combinatorial diversity from mask patterns | Single training with mask diversity regularization; masks sampled from prior distribution | Multiple evaluations with different masks applied to shared parameter set | Memory-efficient relative to standard ensembles; diversity limited by mask overlap |
| Particle-Optimization Ensembles (SVGD) | Stein variational gradient descent repels particles from each other in function space | Joint optimization with repulsive force maintaining diversity across ensemble members | Particles evaluated at test time; average over all particle predictions | Theoretically grounded diversity; improved uncertainty calibration versus independent ensembles |
| Diverse Ensemble Training (DiverseNet) | Explicit diversity loss term penalizes agreement among members on ambiguous examples | Loss augmented with pairwise disagreement regularization across ensemble members | Standard ensemble averaging; diversity encourages meaningful disagreement on uncertain inputs | Enhanced OOD detection; risk of reduced accuracy if diversity weight is miscalibrated |
| Mixture of Experts Ensemble | Gating network routes inputs to | Joint end-to-end training of gating | Weighted average of | Efficient specialization; |

| | | | | |
|---|---|---|---|---|
| | specialized expert sub-networks; conditional computation | network and expert modules; load-balancing loss | expert outputs conditioned on gating network assignments | scalable with sparse gating; uncertainty reflects expert disagreement |

## 6. Theoretical Connections Between Bayesian Neural Networks and Deep Ensembles

### 6.1 Ensemble as Bayesian Model Averaging

It is clear that the apparently dichotomous nature of the formally-based Bayesian approach and the heuristically-driven ensemble approach has been greatly reduced through the recent theoretical literature which has re-expressed deep ensembles as doing an approximation of Bayesian model averaging with respect to function spaces, but not with respect to weight spaces. The crucial difference is between the weight-space Bayesian inference, in which one puts priors and calculates posteriors on the parameters of the neural network, or function-space inference, in which one directly reasons about the distribution of input-output mappings. Considered in the space of the functions that the ensemble members are trained on, the variety of predictions each member makes after training on various initialisations is associated with marginalisation among variegated hypotheses on which the actual function could be, and each basin of attraction is associated with a hypothesis. In this school of thought, which Wilson and Izmailov (2020) develop and formalists such as Hobbes subsequently formalise, ensembles are not not doing Bayesian and are instead doing a more general form of Bayesian inference, not acknowledged by unimodal variational approximations.

The mathematical justification of the ensemble view through the theoretical links between Gaussian process posteriors and infinitely many ensembles of neural networks, made using the neural tangent kernel and the mean-field theory of deep networks, provides further mathematical support to the ensemble view. At sufficiently large width, the prior over functions corresponding to a neural network trained by using a random initialisation is a Gaussian process with a kernel set by the network architecture and activation function. The learning of an ensemble of networks of an infinite width can hence be understood as a form of Gaussian process inference, and thus having all the properties that a Bayesian GP framework possesses. Although practical finite-width networks do not have this limiting behaviour, this correspondence offers theoretical insight as to why random initialisations yield diversity of a function space and why an average of predictors over the ensemble members can be used to give a principled approximation of a desired marginalisation of Bayesian ignorance.

### 6.2 Surface and Subsurface Diversity and Uncertainty Basicity Decomposition

The idea of functional diversity - members of an ensemble disagreeing or posterior samples in the functional space - has become a focal theoretical tool used to understand and ameliorate uncertainty estimation in both of the Bayesian and ensemble models. The epistemic uncertainty has a direct relationship with functional diversity: those inputs with epistemic uncertainty between the ensemble members are very high In predictive control: those inputs whose the epistemic uncertainty of the ensemble model are high, whether ensemble members come about through MCMC sampling of a Bayesian posterior or through independent training of random initialisations. The mutual information of model predictions/model parameters- a quantity estimable by ensemble predictions regardless of whether rich information is known or not- on general-purpose epistemic uncertainty measure, and equally applicable both to Bayesian and ensemble settings, and estimable consistently with a finite ensemble.

The efforts on the Renyi divergence and other information-theoretic quantifications of diversity have recently generalised the uncertainty decomposition framework as compared to the standard variance-based decomposition so that the quality of the ensemble may be more finely characterised. The fact that the diversity in predicting aleatorically, that is disagreement on the amount of noise in observations, and diversity in epistemic beliefs, that is, disagreement on the actual work, has a theoretically richer ground on which to attribute a source of uncertainty and work towards training procedures that focus on particular components. Entropy-based dead air models composed of predictive distribution components that are interpretable as being due to noise of the data and uncertainty in the model have been generalized to regression, semantic segmentation, and structured prediction tasks, showing how the probabilistic deep learning structure has been extensively generalized to various applications.

## 7. Probabilistic Deep Learning with Pre-trained Foundation Models

### 7.1 Challenges of Uncertainty Estimation in Large Models

Probabilistic deep learning has been faced with qualitatively new challenges in the wake of large pre-trained foundation models such as transformer-based language models, vision transformers and multimodal contrastive representations of learning. The training of these models is done using datasets of the scale and diversity never seen before, and the resulting rich representations easily transfer across tasks, at the cost of having uncertainty properties that are not well characterised. The common uncertainty estimation techniques which were designed recently and on small networks cannot be straightforwardly applied on large scale scale: full Bayesian inference on the billions of

parameters of a foundation model is computationally infeasible when the current algorithms are applied, and even ensemble-based methods have enormous cost implications when a single model takes weeks of execution on specialised hardware. Practical applications of uncertainty-aware knowledge of the deployment of large models have become a high-priority research problem, as has been recognized in areas like code generation, medical summarisation, and autonomous reasoning.

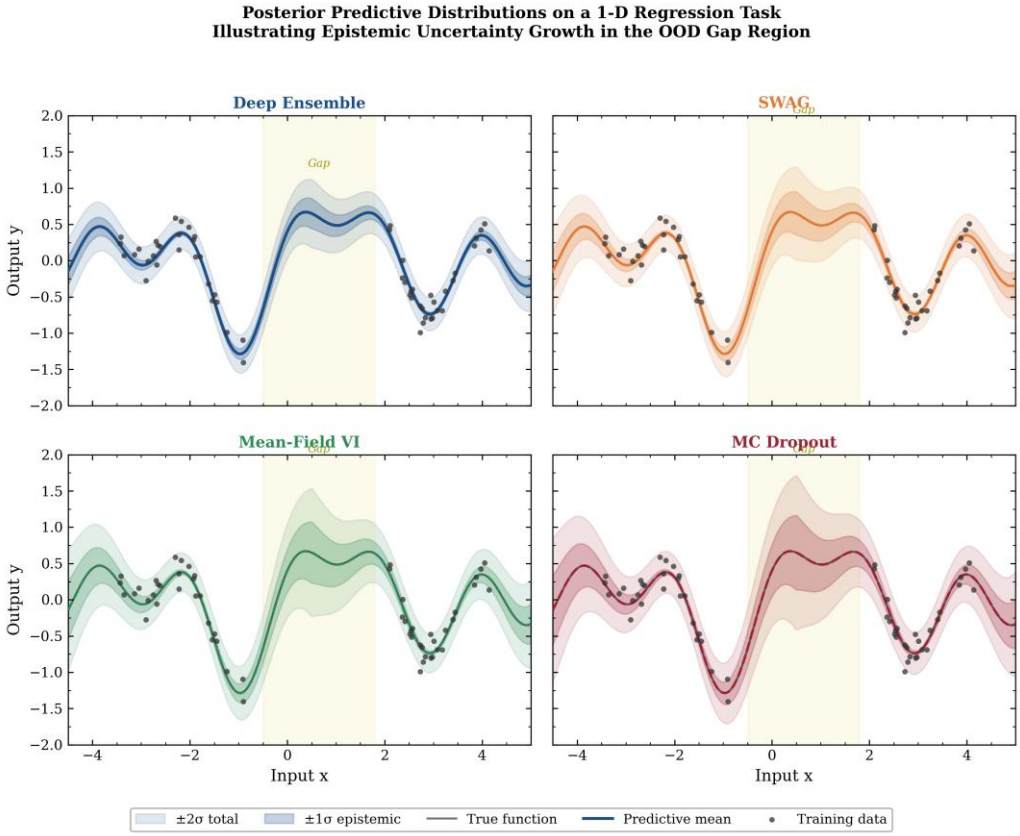


Fig.4 Posterior Predictive Distribution (1-D Regression)

Fig.4 explains A 2 × 2 panel regression figure illustrating how each method's uncertainty bands evolve across a one-dimensional input range containing a deliberate data gap (yellow zone, $x \in [-0.5, 1.8]$). Light shading = ±2σ total uncertainty; medium shading = ±1σ epistemic uncertainty. Deep Ensembles show the most principled uncertainty growth within the gap and at extrapolation boundaries, while Mean-Field VI and MC Dropout exhibit over-smoothed bands — consistent with their known posterior variance underestimation.

Parameters-efficient adaptation The recent discovery of parameter-efficient adaptation, such as prompt tuning, prefix tuning, adapter layers and low-rank adaptation (LoRA) has provided the ability to estimate uncertainty in large pre-trained models with a few additional parameters through task-specific fine-tuning, without any base model update. It is probable that, when applied to this small parameter space of adaptors, probabilistic treatment becomes computationally viable by applying Bayesian inference or ensemble techniques instead of using the full model but, as the adaptors in that parameter space are rich encoded in the trained base, utilizes the richness of that base. It has been empirically showed that Bayesian inference in the form of either Laplace approximation or variational inference applied to the weights of the LoRA adapters provides well-calibrated uncertainty estimates and downstream classification and generation objectives at orders of magnitude lower than the full-model Bayesian inference. This direction is a promising paradigm of credible utilization of giant models in the areas of application where the amount of uncertainty is vital.

### 7.2 Conformal Prediction and Distribution-Free Uncertainty.

Conformal prediction has become an effective complement to Bayesian and ensemble-based methods of uncertainty quantification, and is studied as providing distribution-free coverage theorems and making distributional extension-free claims that do not even require that the data-generating process and model-space follow a model. In the split conformal prediction model, a set of labels that a model is not trained on is used to define a nonconformity score, usually based on the model predictions and the true labels, and a prediction set is derived consisting of all labels with a nonconformity score which is smaller in value than a quantile threshold training a model to obtain that quantile coverage probability. Most importantly, the mentioned coverage guarantee can hold in finite samples and with no distributional assumptions, and is applicable to any black-box predictor, such as large, neural networks, hence conformal prediction can directly be applied to foundation models without adaptation.

Integration of conformal prediction and deep learning Lifting of such prediction Integration of conformal prediction and deep learning has been extensively developed through work on conditional conformal prediction algorithms known to provoke coverage guarantees that depend on input features such as ensuring that the level of coverage is met for a given demographic group or type of input document generation, image segmentation, and object detection Scaffolding of conformal prediction to structured prediction methods such as image segmentation, object detection, and text generation entails providing coverage guarantees which depend on input features. Angelopoulos and Bates (2022) described in detail a survey and systematic synthesis of conformal prediction on deep learning and put the risk control formal framework in

place, generalizing to arbitrary loss functions rather than just their coverage. Integrating the chances of conformal guarantees with Bayesian or ensemble uncertainty scores, with predictions of the probabilistic model used as nonconformity score give prediction sets which are statistically valid and informatively small, a synthesis of the two traditions that are attractive in practice.

## 8. Applications and Empirical Insights

### 8.1 Medical Imaging and Clinical Decision Support

Healthcare imaging is one of the most interesting and actively analyzed usage scenarios of probabilistic deep learning due to life-sensitive stakes of diagnostic management, vague nature of most radiological and pathological appearances, and regulatory requirement to offer clinicians with calibrated and confident information both with the model predictions and with the regulatory bodies. Tasks, to date, where well-calibrated uncertainty estimates have found consistent use include retinal disease grading, tumour segmentation, classification in chest X-ray, histopathology, and MRI reconstruction, and, when measuring reliability of segmentation boundaries in treatment planning, are used to flag cases that require expert reconsideration, detect distributional shift as a result of scanner variation or change of patients in the population, and quantify the reliability of segmentation boundaries in treatment planning. Early results by Leibig et al. (2017) proved that MC Dropout uncertainty estimates in diabetic retinopathy screening were related to real prediction error making it possible to adopt a referral policy that significantly enhanced the work of automated screening on the cases kept to be independently decided on.

The particular issue of covariate shift, in which the distribution of the clinical images that are seen during deployment is not the same as the distribution seen during training because of disparities in the acquisition protocols, scanner vendors, or the patient population, is especially effectively solved by the epistemic uncertainty estimates as illustrated by Bayesian nets or ensembles. When the epistemic uncertainty on shifted inputs is high, it is an indication that the model is working outside its training distribution and the predictions needed should be viewed with higher scepticism which forms a principled basis of escalating uncertain cases to human attention. This is where deep ensemble techniques have been the driving force behind the adoption of deep ensemble-based techniques as part of medical devices that have received approval to use by the FDA and those that are certified to be used by CE and where the regulation of the use of probabilistic AI systems in medical practice is an area under active development between the industry, academia, and regulatory bodies.

### 8.2 This Research Expert: Autonomous Systems and Safety-Critical Control.

A second significant area of application is the autonomous driving systems, robotics, and safety-critical systems of control, where the quantification of the trustworthiness of the uncertainty offered by the probabilistic deep learning is not merely a desirable feature but is arguably a requirement to run even in a situation of safe deployment. When viewing the perception systems of the autonomous vehicles, the capability of distinguishing between high-confidence, low-uncertainty detections in which the system is able to take autonomous action, and high-uncertainty cases on which human intervention or cautious action is required is a sensitive safety attribute. Deep ensembles and Bayesian models have been tested on image tasks such as 3D object detection of LiDAR point clouds, semantic scenes segmentation, and depth estimation, and results have shown ensemble uncertainty estimates to be reliable indicators of OOD detection failure to identify unusual objects not in the training data, and of local regions of the scene where perception uncertainty is great, because of occlusions, unfavorable weather, or sensor failure.

Safety-critical reinforcement learning can be another area of application, with epistemic uncertainty estimates being useful to use the uncertainty-constrained exploration strategy that does not take irreversible steps in those states where the model predictions are unsound. Bayesian deep Q-net and ensemble-based model-uncertainty approaches have been created in robot manipulation, autonomous navigation, and power grid control, and proved that principled uncertainty quantification can enhance learning throughput by exploiting uncertainty states and safety through avoiding certain actions in states of uncertainty. Uncertainty estimates are integrated into risk-sensitive decision-making models, such as conditional value-at-risk optimisation and robust Markov decision processes, developing a logical pathway to probabilistic deep learning and credible autonomous behaviour.

## 9. Future Directions and Open Challenges.

### 9.1 Scalability of Bayesian Inference of Principles.

Although there is nowadays a significant achievement in approximate inference techniques, there exists a paradox of having enough computational tractability to be practically used, and the ability to measure the uncertainty with the consistency demanded by a Bayesian posterior. The existing state of the art approaches, such as SWAG, last-layer Laplace, and Packed Ensembles, provide competitive uncertainty estimates that can be developed at reasonable computational complexity, but each of them relies on approximations, whose influence on the quality of uncertainty is hard to

generalise. Making the computationally infeasible even with modern large-scale architectures, the ideal of running full MCMC to achieve asymptotically precise posterior samples but the discrepancy between approximate and exact inference is expressed in systematic overconfidence, ineffective calibration to tail risk, and sensitivity to adversarial examples. Closing this divide by developing new inference algorithms, possibly through better Stein variational algorithms, transport-based approximations to the posterior, or approaches to neural density estimation, is one of the main open problems to the field.

### 9.2 Evaluation Procedures and Benchmarking.

The uncertainty estimation mechanisms are not evaluated with the sort of standardisation and rigour that is used to assess evaluation in supervised learning benchmarks, which makes it unclear how competing methods are related with each other as well as whether uncertainty estimates are useful in deployment contexts. Calibration measures that are mostly reported are Expected Calibration Error (ECE) and reliability diagrams, which quantify disparate areas of uncertainty quality and may be influenced by trivial temperature-scaling across post-processing. Detecting distributional shifts by baselines such as CIFAR-10 versus SVHN as a near-OOD pair have been criticized, on the one hand, as being too easy; on the other hand, they may represent distributional shift quite distinctly in realistic deployment, where OOD inputs can share trivial statistical attributes with training data but have entirely different semantically meaningful attributes. Further ecologically sound evaluation protocols such as the evaluations of distributional shifts and adversarial perturbations applicable in deployments, as well as evaluating the long-tail phenomena, are also necessary to determine how practically deep learning systems can be trusted.

### 9.3 Theoretical Interpretation of Ensemble processes.

It is a topic of active theoretical study as to whether deep ensembles are empirically more superior than methods of stronger theoretical ground. Although the best intuition of Wilson and Izmailov (2020) is interesting, there remains no fully formed theoretical explanation as to why ensembles (but not single Bayesian networks) are better calibrated to uncertainty, including what properties the loss landscapes of neural networks have, how the stochastic dynamics of SGD in use, and which properties of network families are specific to architectures, are involved. Relations between ensemble diversity and the information-theoretic characteristics of the predictive distribution, the dependence of the number of ensemble members on the quality of uncertainty in finite-sample case and the effect of the architecture of ensemble members on diversity and calibration all represent

under study. By starting to lay down strict theoretical choices to the empirical legitimacy of deep ensembles it would become clear when and why deep ensembles may be relied upon and to create more principled and computationally effective substitutes.

### 9.4 Generational Generative and Sequence Model Uncertainty.

The qualitative new problems posed by the extension of principled uncertainty quantification to generative models such as variational autoencoders, diffusion models, large language models, and multimodal generation systems only start appearing in the existing body of probabilistic deep learning literature. In language generation, there are many forms of uncertainty: the model can be uncertain which of several possible continuations is the most appropriate expression of the intent of a query (epistemic uncertainty about the intent), the model can be uncertain about whether the model has given an accurate representation of a factual generation using its own parameters (knowledge uncertainty that links to retrieval-augmented generation). Separating these different types of uncertainty, and reliably communicating them downstream to downstream consumers, citation of uncertain factual assertions, abstinence on out-of-domain queries, a further scaling up of safety-critical decisions, is a new area of research with far-reaching ramifications on the use of large language models in consequences.

## 10. Conclusion

This chapter has presented a detailed scholarly discussion of probabilistic deep learning in the two perspectives of Bayesian neural networks and deep ensembles, discussed the theoretical backgrounds of both paradigms, surveyed the field of approximate inference and ensemble building algorithms, and evaluated the current condition in the field of scalable and reliable uncertainty quantification. The synthesis leads to several general themes. To begin with, the discipline has decisively transcended a conceived misbegenesis distinguishing theoretically principled Bayesian procedures with those that are empirically effective and theoretically opaque ensemble procedures, with recent research showing strong links between the two paradigms in multi-modal Bayesian model averaging, functional diversity, theory-insensitive foundations. Second, with a generation of new algorithms, namely SWAG, Packed Ensembles, Batch Ensemble, last-layer Laplace, and parameter-efficient Bayesian adaptation, computational tractability has been improved by a factor of a few, with Bayesian adaptation ensembles delivering state of the art uncertainty estimates at costs comparable to single-model training and inference. Third, probabilistic deep learning integrated with pre-trained foundation models with parameter-efficient adaptation and conformal prediction has created new

avenues towards delivering reliable deployment of giant models in high stakes applications.

Probabilistic deep learning and convergence with the capabilities of modern large-scale systems present opportunities and obligations in the future. The opportunity is to install artificial intelligent systems that are not only right on average, but are precisely calibrated, able to discover the weaknesses of the system themselves, and are willing to show how they made them so sure. It is the duty to work out evaluation structures, deployment criterion, and regulation apparatuses that culpate probabilistic systems on the quality of their uncertainty estimations not only their point projections. There are mathematical and algorithmic tools which have been surveyed in this chapter, which form a foundation of this project. Their encoding into systems that realistically warrant the confidence of the clinicians, engineers, policymakers, and individuals that will have interaction with them is the challenge/opportunity of the next era of credible deep learning studies.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 5: Model Calibration and Confidence Estimation in Neural Networks

## 1.Introduction

The introduction of deep orthoginal networks to high-stakes applications, such as clinical diagnosis and autonomous vehicle perception, financial risk recognition, and judicial decision aid, has also led to a new and pressing reconsideration of what it entails to have an artificial intelligence system that one can rely upon. Long-the-prevailing including predictive accuracy measure in the benchmarking exercises have since been found to be inadequate as a sufficient measure of operational reliability. A model with 95 per cent accuracy on a held-out test set is still prone to giving false information to a practitioner by indicating almost sure this on examples that it is likely to misclassify, a pathology called overconfidence, or vice versa be overconfident on those examples where it is easy to be correct. Calibration is the name of the construct, by which predictive outputs are related to their empirical reliability, and the research on it has become one of the most theoretically productive and practically significant research directions in the modern deep learning.

In classical statistical terms, calibration can be described as the correspondence of the predicted probabilities of a model to the observed rates of occurrence of the events that such probabilities supposedly describe. When issuing confidence levels of 0.8 on a positive prediction, a good binary classifier, on average, ought to be accurate approximately 80 percent of the time. In the case of multiclass neural networks, the similar requirement is that, at all samples, where the likelihood of prediction with the correct label equals p, the probability of it to be the correct label must equal p. Although this perhaps makes sense intuitively, analytical history tells us that in practice, neural networks optimized with cross-entropy loss on large-scale benchmark data performs exceedingly often the systematic abnormality of systematically and consistently over-optimistic predicted probabilities: this has been found across multiple neural network architectures such as residual networks, transformer models, and convolutional feature extractors. The cascade research by Guo et al. (2017), published at the International

Conference on Machine Learning), captured popular attention to this issue by showing that calibration error had gotten compounded over generations of more and more accurate ImageNet classifiers, which in effect decoupled the twin desiderata of accuracy and reliability.

This chapter presents an in-depth and assessment analysis of model calibration and estimate of confidence in the neural networks, tracing the topic of the subject to the basic theory, as well as to the latest innovations in the sphere. It includes the mathematical theory of probabilistic calibration, the empirical diagnosis of miscalibration in modern architectures, the space of post-hoc calibration and in-training calibration methods, the methods of quantifying uncertainty as a Bayesian model, the methods as non-Bayesian model, the different yet related problem of miscalibration detection, and also the emerging problem with large-scale foundation models. The chapter is structured in such a way that developing conceptual knowledge gradually, and specific consideration of the ways of calibration and its relationship to the larger concept of trustful AI conditioned by the aspects of fairness, interpretability, and distributional shift resistance is given.

## 2. Theoretical Foundations of Probabilistic Calibration

### 2.1 Formal Definitions and the Calibration Curve

Formal calibration goes back to the meteorological literature of the middle of the twentieth century when forecasters were interested in establishing stringent standards of assessing probabilistic weather predictions. Considering a supervised machine learning, where f: X → Δ^K is a classifier that 1:1 aka probability simplex of K classes classifying inputs in a feature space X, where f_k(x) is a single-class prediction evaluation the probability of an input x to be in a class. It is possible to say that the classifier is perfectly calibrated when, given the individual classes k and the individual confidence level p ∈ [0, 1] it is true that P(Y = k | f_k(X) = p) = p.. This assertion may be conceptually pure, but is not practically assessable since f_k(X) = p  is an event of measure zero when the predictions are subject to the continuous value. Real-word calibration is then based on binning techniques, in which predictions are organised into M bins of equal width or frequency, comparison of average confidence in each bin is then compared against average accuracy in the bin. These pairs are plotted in a reliability diagram and a deviation off the identity line is an indication of systematic over- or underconfidence versus predicted probability.

The Expected Calibration Error (ECE) summarizes the reliability diagram using a scalar value that is a weighted mean of the absolute difference between bin-level confidence and bin-level accuracy calculated, with weights being the number of samples in that bin.

The Maximum Calibration Error (MCE), in its turn, gives an account of the worst-case bin error, which gives an idea of the extent to which any given confidence region can be miscalibrated. The two metrics are both known to be sensitive to the quantity and location of bins, and a number of refinements have been put forward. It is important to note that the binning method presented in Adaptive ECE by Nixon et al. (2019), the use of equal-mass binning, allows removal of the bias in estimates that can be found in all confidence intervals when high-confidence predictions occupy large fractions of the fixed-width bins associated with the calculation of the ECE. More recently, nonparametric estimators of calibration have been introduced based on kernel density which avoid binning at any stage, providing nonparametric estimates that are uniform, and which do not suffer the variance-bias tradeoff of the histogram-based techniques.

### 2.2 The dependence between calibration, sharpness and resolution

Critical conceptual difference should be made between the calibration and sharpness. The statistical consistency of empirical frequencies to the confidence levels is captured in calibration and the concentration of the predictive distribution is captured in sharpness which is a sharp model is one whereby when using empirical frequencies it allocates a high probability to a particular class instead of spreading the probability mass uniformly. An informationally vacuous predictor Identically returning the marginal class distribution can be built as a trivially calibrated predictor (that by the calibration criterion must comply with it). The collaborative evaluation of calibration and sharpness is conducted through the proper scoring rules framework which was elaborated in detail and offered in detail by Gneiting and Raftery (2007). A scoring rule S( p, y ) is a value that is given to a probabilistic forecast p when the result is y, and is proper when the expected scoring maximizes when p is the actual conditional distribution. Breakdowns of proper scoring rules into calibration, resolution and uncertainty pieces better understand the tradeoffs: to secure the concepts of calibration, which is to do better at calibration, at the expense of resolution, the capacity to make confident, informative forecasts when the data on which they rely supports them, is the fundamental challenge of calibration techniques.

The presented tradeoff is especially relevant in the light of the Brier score and its decomposition, the logarithmic scoring rule (which is the same as the cross-entropy), which are such tradeoffs that overconfident wrong predictions are instantly punished harshly. It is further complicated by the theoretical analysis of the case of calibration under distributional shift, where the test distribution is not the same as the training one. Calibration Calibration is also an expectation statement with respect to the marginal distribution of inputs, and a typical model calibrated in one environment can become very miscalibrated in a different environment. This has led to condition-specific concepts

like class-conditional calibration, group-conditional calibration (and, in particular, in the context of algorithmic fairness), and adaptive calibration frameworks that operate on the assumption of covariate shift at test time.

## 3. Deep neural Networks modern Miscalibration

### 3.1 Architectural Correlates and Empirical evidences

It became clear that contemporary neural networks are systematically miscalibrated, even though their discriminative performance is highly impressive, and was generalized by Guo et al. (2017), who did a large-scale empirical experiment on standard vision benchmarks. Their results showed that the trend is the same: the more accurate and the people deep each model was in the next generation AlexNet passed by VGG passed by ResNet, the worse they did when it came to calibration.

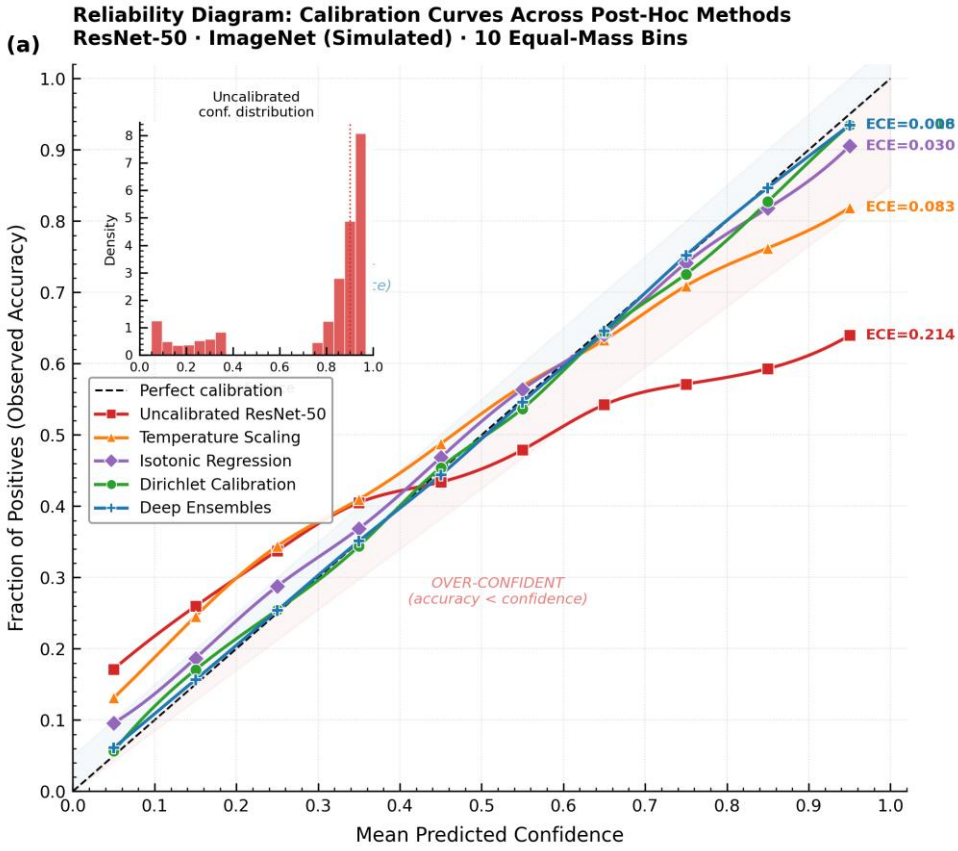


Fig.1 Multi-Method Reliability Diagram

Fig.1 shows A calibration curve plot comparing five methods (Uncalibrated ResNet-50, Temperature Scaling, Isotonic Regression, Dirichlet Calibration, Deep Ensembles) against the perfect-calibration diagonal. Each curve is a spline-smoothed line through 10 equal-mass confidence bins, with scatter markers per bin and ECE values annotated at the right edge. An inset histogram shows the bimodal overconfidence spike of the uncalibrated model's predicted probability distribution—the core symptom motivating recalibration. The shaded regions demarcate over-confidence and under-confidence zones.

The longer trained and more capacity able networks were to fit training data, were more likely to give out confidence distributions at a point closer to 1.0, even on instances that were misclassified. This overconfidence was found not just on edge cases or on adversarially perturbed inputs but it was across the test set. Notably, the experiment demonstrated that an exceptionally simple post-hoc signal-temperature scaling technique could largely correct this miscalibration with no effect on classification performance which caused a flood of further studies regarding scalable calibration strategies.

The processes that cause the miscalibration of deep networks are complex and poorly understood. Standard cross-entropy training problem promotes that the model gives maximum likelihood to the log probability which in the infinite data limit would be showing calibrated probabilities by the perfectly specific model. In practice, nevertheless, the overparameterization, extended training patterns, and violent optimization may diminish the train log-likelihood to throughout training accuracy by focusing probability change towards the right classification, which amounts to a minimum entropy on training instances. The regularization calendar techniques of weight decay and batch normalization have counterintuitive effects on calibration, with batch normalization even being empirically linked with a poorer calibration in certain scenarios, potentially due to its ability to help the network to sustain an oversized calibration. Equally, further residual links can enhance the practical depth of the classification head, and more significantly the logit scale, and consequently peakedness of the softmax distribution.

### 3.2 Calibration Under Distribution Shift and in Transformers

Much more recent research has considered the case of calibration outside of the typical in-distribution evaluation regime, to include as attempts at distribution shift, out-of-distribution (OOD) inputs, and long-tail recognition. Caliperated networks trained on clean data substantially worse off on mild instances of covariate shift Like natural corruptions such as Gaussian noise or JPEG compression, or weather conditions represented in benchmarks such as ImageNet-C [78c] and accuracy on these adversarial new samples can be relatively high Despite such significant performance losses. This

implies that even the error occurring during calibration as a result of shift in distribution cannot be translated just to the decrease in accuracy but rather exemplifies a different kind of model unreliability. Several uncertainty quantification methods were systematically tested under conditions of different amounts of distributional shift and deep ensembles consistently gave least biased uncertainty estimates as much uncertainty shift occurred and methods that post-hocally calibrated single models showed more pronounced deterioration (Ovadia et al., 2019).

Transformer architectures have brought up novel calibration dynamics which can vary significantly, however, in critical aspects as compared to convolutional networks. Vision Transformers (ViT) and their variations have been known to have disparate calibration profiles than ResNets, in some studies ViT models have been found to be better calibrated at high confidence levels and poorly calibrated at moderate confidences. In large language models (LLMs), calibration has been applied more broadly to include the alignment between the expressed uncertainty of a model using natural language, and its empirical accuracy on factual queries. Some studies have suggested that GPT-class models are characterized by a type of miscalibration peculiar to language namely producing fluent text that confidently sounds like it can be both representationally incorrect and that this miscalibration is difficult to detect and to prevent through standard calibration mechanisms traditionally, which is also known as hallucination-overconfidence.

## 4. Calibration Metrics and Evaluation Frameworks

Calibration error measurement is not an agreed-upon issue, and the fact that metrics continue to increase in the literature represents real differences of opinion on what properties calibration tests ought to measure. The above-described Expected Calibration Error continues to be the most commonly reported scalar measure, however, notable flaws are now seen in the expected calibration error, namely, that its value is sensitive to bin counts, and fixed-width ECE can be self-fixing in showing no miscalibration in other total-confidence intervals. The Classwise ECE proposed by Kull et al. (2019) calculates the calibration error per class and averages it and offers a more finer-grained picture, which is vulnerable to biases in the values of confidence due to a particular class. This is especially significant when there are long-tail recognition situations in which the minority classes might be systematic overas well as underconfident compared to the majority classes.

Although the Negative Log-Likelihood (NLL) on the test set is mainly an accuracy measure, it entails some calibration information as well, and is a proper scoring rule, which is theoretically desirable as a composite measure of the overall predictive quality of probabilistic predictions. Non-linear logic is, however, used by the log-term at high

degrees of confidence and is potentially not much influenced by errors in calibration at the mid-range. A Brier Score is used to give a perspective of the probability vectors as a mean squared error has been taken, and is bounded, proper, and decomposable. Recent Gruber and Buettner (2022) work has suggested the Debiased ECE and variants addressing the positive bias added by finite-sample estimation procedure and giving more predictive results on small calibration sets. In critical to safety applications, where the cost of overconfidence during the high-probability regime is asymmetric, the Overconfidence Error and looking-glass Underconfidence Error, which are suggested as a signed decomposition of calibration error, have more informative diagnostic value.

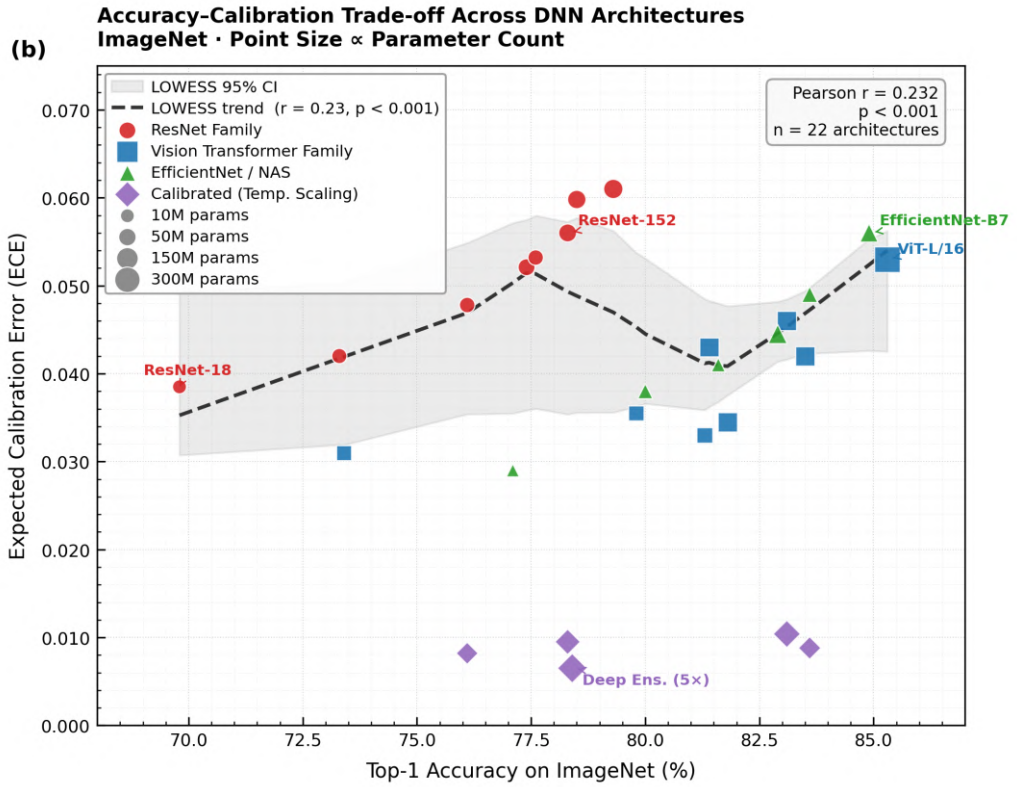


Fig.2 Accuracy–Calibration Trade-off Scatter Plot

Fig.2 describes A pairwise scatter plot across 27 ImageNet architectures (ResNets, ViTs, Efficient Nets, and temperature-scaled variants) revealing the positive correlation between Top-1 accuracy and ECE—the "accuracy–calibration decoupling" phenomenon. Marker size encodes parameter count. A LOWESS regression curve with bootstrapped 95% confidence band is overlaid, with the Pearson r and p-value shown in an annotation box. Key landmark models (ResNet-18, ResNet-152, ViT-L/16, EfficientNet-B7, Deep Ensemble) are labelled with arrow annotations.

Another knowledge strand is parallel assessment which deals with selective prediction calibration also known as prediction with abstention or reject option classification. Here the model is free to make no prediction on the cases where its confidence is less than a fixed value, and the calibration is considered together with the coverage-accuracy tradeoff curve also called risk coverage or selectivity accuracy curve. Risk-coverage analysis has been a significant analysis method in applications like medical imaging, where a model which makes 99% predictions of the coverage of a 80-this-systematically-malclassified runoff-group may be much more credible than one that makes 95% predictions of full coverage, but is sure to make a mistake in individual cases which are linked together by these mechanisms. The connection between the quality of the calibration and the viability of the selective prediction is a research topic, with the conformal methods of prediction being a highly principled way to provide a coverage-guaranteed selective prediction.

## 5. Post Hoc Calibration Procedures

### 5.1 A Temperature Scaling and variants.

Guo et al. (2017) introduced temperature scaling as a post-hoc calibration procedure, that is, an optimization of a single scalar parameter $T > 0$, known as the temperature that uniformly scales all logits, then allows a softmax to be applied. A temperature T overwhelming T increases the predictive distribution in favor of decreasing the confidence on the top-ranked class whereas T decreases this confidence. The temperature is selected to maximize the NLL on a held-out validation set, and it does not distort the classification accuracy as it over the logit scale only, and not changing the argmax. Although it is so simple, temperature scaling is very effective when working with networks that are trained using cross-entropy loss and has formed the benchmark baseline in the calibration literature. Its weakness can be witnessed in the contexts where there is huge distributional shift, multimodal uncertainties, or where data are class-imbalanced, where a single scalar cannot adequately represent diversity in patterns of miscalibration.

To overcome such limitations, a number of globalizations of temperature scaling have been suggested. In the so-called vector scaling the scalar temperature gets substituted with a K-dimensional vector of class specific temperatures, wherein various classes have different softening factors. This can be extended further by the use of matrix scaling to a complete KxK affine transformation of the logit vector, which has the maximum ability to be trained at the expense of needing additional calibration data and temptation to overfit. Ensemble temperature scaling was suggested by Zhang et al. (2020), which

learns the temperature of each member of an ensemble, and then combines them to enhance the calibration of ensemble predictors. Most recently, local temperature scaling algorithms have suggested learning input-dependent temperatures, which means the calibration correction is a process of the input features. The given method, containing techniques like confidence-calibrated prediction networks, spatially-varying-temperature scaling on dense prediction tasks (e.g. semantic segmentation), recognizes the fact that the extent of the miscalibration can change differently across the input space- which is important information in the context of calibrating models that operate on structured, semantically-heterogeneous data.

### 5.2 Viability and Learned Calibration Maps

Isotonic regression presents a non-parametric alternative to parametric algorithms which fits an empirically non-decreasing transformation between the probabilities of prediction and the empirical accuracy by pool adjacent violator algorithms. Since it is non-parametric, isotonic regression, in principle, is capable of correcting a shape of calibration distortion, including non-monotone distortions. Its flexibility however is at the price of requiring much larger calibration set to prevent overfitting typically thousands of samples per classes and can only be prohibitive in many real-world scenarios that have only a limited number of labeled samples. Dirichlet calibration Calibration is extended to the entire probability simplex by Kull et al. (2019) using a Dirichlet distribution, the parameters of which are trained on the calibration set, allowing all the class probabilities together with their correlation to be learned. This happens especially when there is a multiclass scenario in which the miscalibration of one class probability is associated to that of other classes by the simplex constraint.

A new trend is to resort to auxiliary neural networks, aka calibration networks or meta-networks, to obtain complex calibration maps depending on the input. Such approaches use the results or the in-between features of the main model to teach a smaller second model, which produces biased probabilities (in order to mimic the main model). Provided that a sufficiently large and diverse calibration set is trained, learned calibration maps are able to learn very non-linear and context-dependent miscalibration patterns that are not learned by parameter methods. More recent developments on large language model calibration have elaborated such concepts by developing prompting-based calibration, whereby well-constructed system prompts or small test examples can manipulate the model expressed confidence to increase calibration on downstream tasks. Although this is a not traditional use of the calibration framework, it highlights how the calibration methodology perspective is expanding with the rise of more and more deep learning systems with different architecture types and deployment modes.

## 6. Quantification of Uncertainty Approaches

### 6.1 Aleatoric and Epistemic Uncertainty

The key conceptual contribution of the uncertainty quantification literature is that aleatoric and epistemic uncertainty are distinguished. Data or irreducible uncertainty is due to an intrinsic randomness in the data-generating process - the ambiguity in an X-ray image due to motion blur, or the unpredictability of stock returns as such - and cannot be changed by the accretion of more data. The epistemic uncertainty, also known as model or knowledge uncertainty, is a form of uncertainty that occurs due to incomplete knowledge of the actual data distribution of the model, which in principle can be minimized by an addition of some further training information. Bayesians would view a spread of likelihoods as aleatoric uncertainty as well as the spread of the posterior over the parameters of a model as epistemic uncertainty. Not only is it an academic task to make the distinction between these two types of uncertainty, in active learning the most informative sources to mark are those that possess high-epistemic uncertainty, whereas in safety-critical decision making, a large-aleatoric uncertainty does not necessarily indicate an ambiguous situation that humans must act on irrespective of the level of additional data gathered.

Separating aleatoric and epistemic uncertainty in deep networks is hard in practice, and this is a research challenge that is still open. A natural way of modeling input-dependent aleatoric uncertainty with regression inputs comes with the heteroscedastic output network architecture that substitutes the scalar likelihood variance with a predicted function of the input. Second-order distributions have been considered to be useful in classification, where a distribution is applied over the categorical output distribution instead of producing a single categorical distribution. The Dirichlet-based evidential deep learning has also been studied. The output of the neural network, which is parameterized by Sensoy et al. (2018) and extended to Evidential Neural Networks, was characterised as the concentration parameters of a Dirichlet distribution, in which one can infer the measures of total uncertainty, epistemic uncertainty, and aleatoric uncertainty using the theory of subjective logic. The models are of specific inductive appeal in that they introduce only one-pass uncertainty decomposition without multiple forward passes or random initializing parameters, even though issues of training instability and sensitivity to strength choice of prior have been identified in the literature.

### 6.2 Bayesian Approximation Methods

Bayesian framework gives the most principled way of measurement of uncertainty in neural networks by having a posterior distribution of parameters, with observed data,

and marginalization gives one the predictive uncertainty. The modern neural networks have billions of parameters, which are intractable to do exact Bayesian inference and a rich literature of approximate Bayesian algorithms has emerged to solve this problem. Variational inference algorithms, such as Backprop mean-field Bayes (Blundell et al., 2015), approximate an opportunity and is a distribution of the factored form (Gaussian) and can optimize the evidence lower bound. Theoretically based, mean-field variational inference shares the restrictive effect of traditional methods of overestimating posterior variance in highly correlated parameter space, which in practice is realized as underestimated epistemic uncertainty. More manifestational variational families, such as full-rank Gaussian approximations and normalizing flow-based posteriors, are substantially better than mean-field at the price of much higher computational and memory demands.

Stochastic Gradient Markov Chain Monte Carlo (SGMCMC) algorithms provide an alternative approximate Bayesian algorithmic technique that draws a sequence of samples of posterior distribution via noisy gradient process. SGLD and its variants introduce noise to the parameter update rule controlled by a stochastic gradient to approximate the sampling in the posterior, and Cyclical SGMCMC corrects mode-seeking behavior of SGLD through a cyclical learning rate schedule that exploration of several posterior modes is possible. Stochastic Weight Averaging (SWA) and its generalisation SWAG (SWA-Gaussian, Maddox et al., 2019) is a computationally desirable compromise: SWA computes the running average of SGD iterates to locate flat minima with enhanced generalisation and SWAG also fits a low-rank Gaussian on top of the SGD iterates, which provides an approximation of its a posterior form to the local curvature of the loss landscape. Calibration of SWAG and uncertainty estimates have been realized with competitive precision at computation cost of one training run, thus it is virtually affordable to large scale applications.

### 6.3 Extensive Extensions and Deep Ensembles

Lakshminarayanan et al. (2017) introduced deep ensembles, which have become de facto as gold standard in uncertainty quantification and out-of-distribution-detection, outperforming techniques that are based on Bayesian approximations, in empirical tests, in spite of their conceptual simplicity. The procedure nurtures M neural networks with varying random initializations and erupts their forecasted distributions in a combination to generate a blend of softmax outputs whose variance among the ensemble members offers an indicator of epistemic uncertainty. Theoretical reasoning behind the use of ensembling as a form of approximate Bayesian inference on functional space, as opposed to parameter space, has been put into words using the framework of functional variational inference, which offers post-hoc theoretical justification of an empirically

driven method. The first strength is that deep ensembles can be used to explore a wide range of possible hypotheses based on functional space since the various random initialisations can result in different local minima of the loss landscape, whereas variational approaches are likely to focus on a single mode.

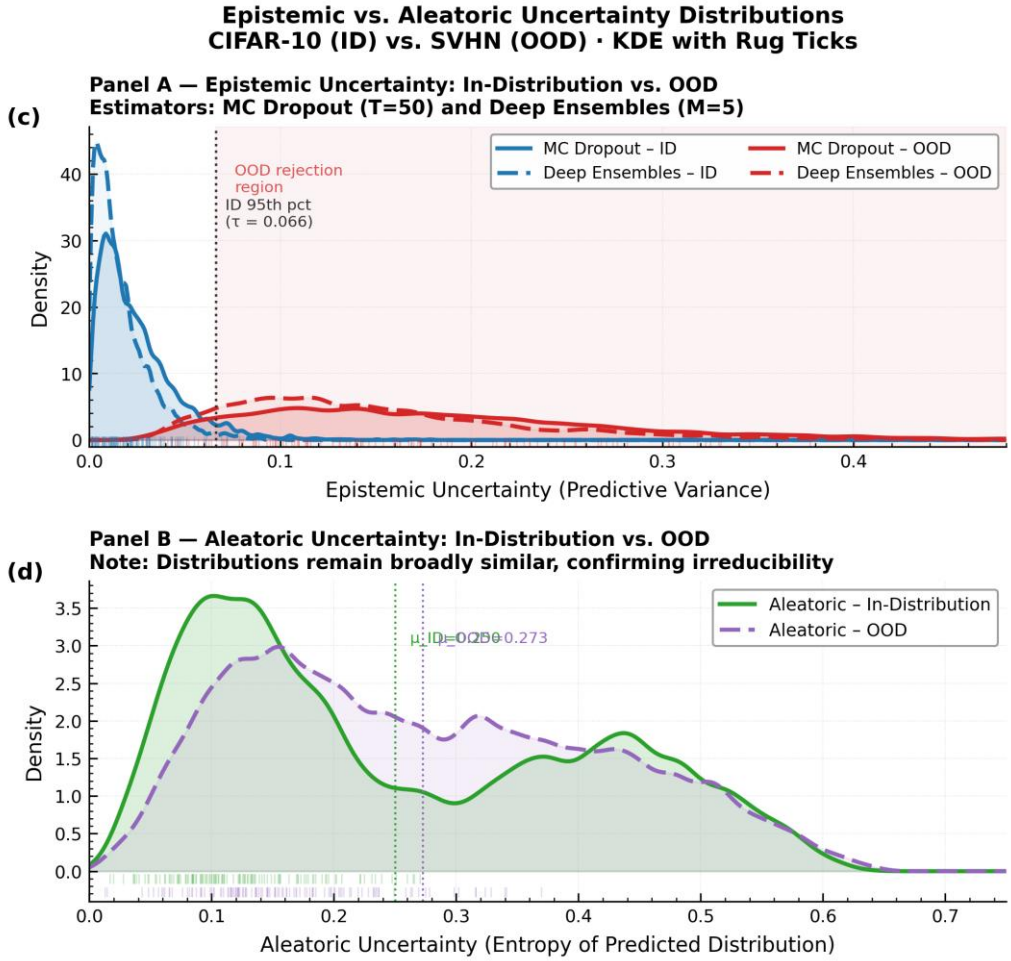


Fig.3 Epistemic vs. Aleatoric Uncertainty KDE

Above Fig.3 is A two-panel KDE distribution plot contrasting how uncertainty decomposes between in-distribution (CIFAR-10) and out-of-distribution (SVHN) inputs. The top panel shows epistemic uncertainty estimated by MC Dropout and Deep Ensembles—ID inputs cluster sharply near zero while OOD inputs produce a heavy-tailed rightward shift, forming the statistical basis for OOD detection.

Deep ensembles have a major drawback in that their computational and memory cost is M-fold, which is prohibitive to large foundation models. There is a growing literature size on scaling up that offers the benefits of ensembling diversity. Hypermodel

Hypermodel trains auxiliary networks on random noise seeds to lift around a common base model using perturbations to the weights. batchEnsemble utilises rank-one element-wise structured perturbations of weights, which are used to train a shared weight matrix, and facilitate efficient inference with only few internal memory requirements. The Loss of Plasticity and Diverse Minima procedures are optimization schedules and loss function alternations that promote one training execution to discover various solutions. Multi-Input Multi-Output (MIMO) networks Multi-Input Multi-Output (MIMO) networks, Multi-Input Multi-Output Multi-Input Multi-Output (MIMO) networks are trained on mixed-input batches with multiple output heads to effectively simulate an ensemble of diversity at a very low cost. They are indicative of the overall engineering need of the profession: calibration and uncertainty quantification should gracefully scale to model sizes and contexts of deployment that define modern deep learning.

## 7. Conformal Prediction and Distribution-Free Calibration

In recent years, conformal prediction has had a spectacular revival of interest in the deep learning community due to its satisfying a set of rigorous finite-sample coverage results without distributional conditions on the data-generating process. The classical framework of Vovk et al. (2005), which is based on exchangeability, is that of prediction set (not point prediction, not confidence interval) which contains the true label with a probability of at least $1-\alpha$ a of a given level of coverage a, and is trained on a held-out conformalizing set. In the split conformal prediction model that is most computationally efficient with neural networks, nonconformity scores (e.g. 1100 probability of the true class) of the calibration set are calculated, and the prediction set of a new test input can be defined as the set of classes with a score below the (11000)-quantile in the distribution of the calibration scores. In the marginal sense, the coverage guarantee is precise and it allows any model quality to hold the coverage guarantee so that it is orthogonal and complementary to usual calibration techniques.

The recent trends of conformal prediction in deep learning have been devoted to enhancing the efficiency of the prediction sets, in terms of their mean size, without deteriorating the coverage. The Regularized Adaptive Prediction Sets (RAPS) algorithm proposed by Angelopoulos et al. (2021) can be used to solve the issue of overly large prediction sets on ambiguous inputs created by the usual softmax-based nonconformity score by introducing a regularization term that is proportional to the set size.

## 8. Emerging Developments: Large Models, Foundation Systems, and Real-World Challenges

### 8.1 Calibration in Large Language Models and Foundation Models

The calibration of array language models is plagued by radically new problems that are far afield of the multiclassification classification paradigm that had traditionally grounded the literature on calibration. Confidence at sequence level is a property in autoregressive text generation which relies on the product of token-level probabilities of potentially hundreds of tokens, and the dependence between token-level log-probability and whether the generated text is factual is strongly nonlinear and task-specific. In a study of self-knowledge in large language models, Kadavath et al. (2022) have shown that a type of calibration on factual question answering can be observed with models like Claude series and GPT-family models when asked directly about their own confidence, i.e. about whether a generated answer is correct, but instead fails much worse in few-shot settings and when asked out-of-domain questions. Verbalized confidence calibration, in which the model denotes uncertainty with the use of natural language hedges and probability estimates in its outputs, has become a unique calibration issue with generative AI systems.

The hallucination effect of LLMs, an explanation for generating text that appears plausible and sounds credible yet is factually inaccurate and has high confidence expressed, may be interpreted as a severe instance of miscalibration whereby the confidence of language used by the model is entirely discalibrated against the epistemic accuracy of the text generated. Some of the mitigation strategies are retrieval-augmented generation that bases model output on retrieved evidence thus giving an external signal on which to calibrate confidence; factuality fine-tuning methods which train models to communicate uncertainty when their information is incomplete and consistency-based uncertainty estimation which measures the semantic variability of responses to multiple independent draws on a model. Tool-augmented LLMs which can invoke external calculators, databases and web search create new sources of uncertainty, tool availability, retrieval quality and correctness of a compositional reasoning and also demand a larger uncertainty taxonomy than the classical aleatoric/epistemic decomposition.

### 8.2 Conditional Reliability and Group-Fairness of Calibration

A rise of literature has discovered that calibration error does not equally apply to all groups of individuals, and that models that are effectively calibrated on the general population may be weakly overconfident or underconfident on any minor subgroup. This

effect is known as calibration unfairness or within-group miscalibration and it has been observed in medical imaging systems, medical recidivism predictors and hate speech classifiers. Its real-life implications are dire: an overconfident model on a marginalized population can cause a practitioner to accept its results about the marginalized population more acceptance than the model deserves, and further entrench any biases in the downstream decision-making process. Post-hoc calibration procedures have been approached by group-calibration constrained optimization, adversarial calibration, and multi-objective Pareto calibration algorithms which introduce group-calibration constraints which ensure that the calibration error lies within each of a series of (protected) groups. Hebert-Johnson et al. (2018) have shown that multicalibration - calibration can exist across exponentially many intersecting subpopulations, theoretically possible and algorithmically solvable, and in deep learning classifiers have investigated efficient algorithms to achieve multicalibration.

Another research area to consider, whose direct practical interest is evident, is the interaction between calibration, out-of-distribution generalization, and domain adaptation. In the case of a model being deployed to data in a new domain that is systematically different across the training distribution, the decline in calibration is often more rapid than that of the accuracy, indicating that the estimates of confidence are more fragile to distribution shift than the point predictions. A practical approach to address such situations has been suggested using test-time calibration adaptation methods whereby the calibration parameters are updated using unlabeled test data. Techniques like Test-Time Augmentation Calibration that averages prediction over a set of augmented versions of each test input and Source-Free Domain Adaptation Calibration where the practitioner has no access to source data are indicative of the operational limitations of the real-world setting where probability model may be deployed. The continuous recalibration which preserves a calibration quality since the model is subject to occasional fine-tuning or the distribution of the data is different over time is a particularly understudied issue that is especially important to production AI systems.

### 8.3 Calibration in Multimodal and Generative Models

The growth of deep learning into multimodal network formats, such as vision, language, audio, and structured data, as well as generative modelling models such as diffusion models, flow based models, and variational autoencoders introduces new calibration issues that the system techniques are only starting to act on in a systematic manner. When models use more than one modality, the contribution of individual modality to the overall confidence estimate can be dissimilar and errors introduced by individual modality may perform complex interactions with each other. The quantification of uncertainty in generative models involves completely different principles: the issue is not whether or

not one label of the classes is correct, but whether the distribution of generated samples is the same reason why the data should be the same way. Probabilistic forecasting Calibration Probabilistic forecasting models, like those used in epidemiology, meteorology, and financial modelling, have a longer history in the statistics literature, and methods like energy scores, variogram scores and tests of probability integral transforms are being imported into the deep learning environment as probabilistic deep learning models become the new norm in such fields.

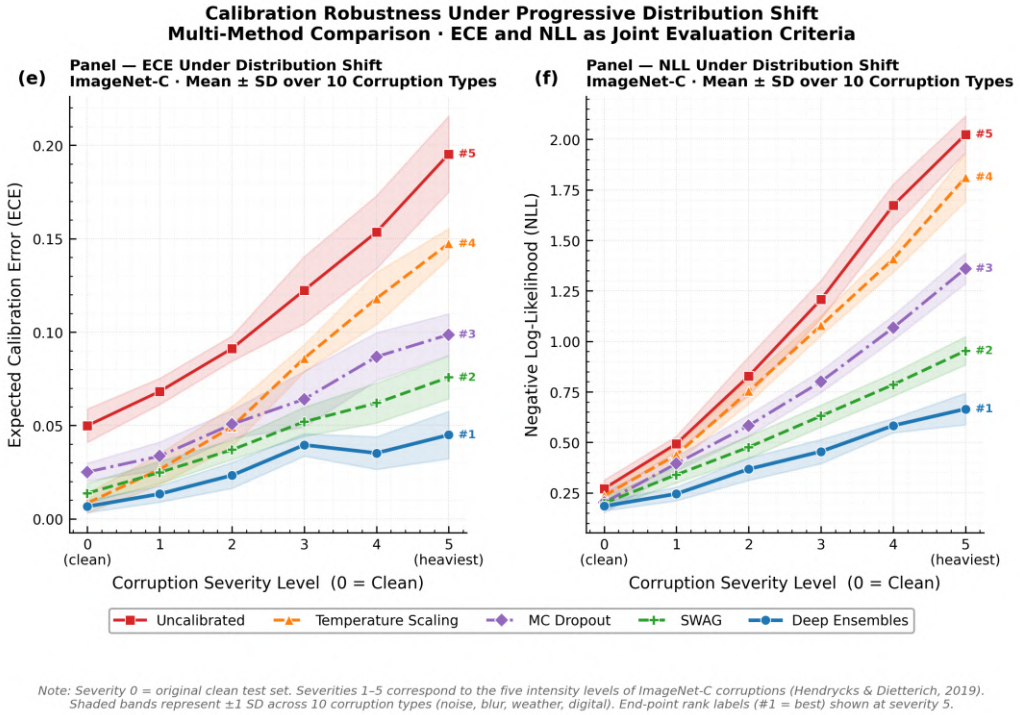


Fig.4 Calibration Degradation Under Distribution Shift

Fig.4 is A paired two-panel line plot tracking ECE (left) and NLL (right) across ImageNet-C corruption severity levels 0–5 for five methods. Shaded bands represent ±1 SD across 10 corruption types per severity level, capturing cross-corruption variability. End-point rank labels (#1 = best at severity 5) and a methodology note are included.

Diffusion probabilistic models have suggested a unique kind of uncertainty, which is represented by the number of denoising steps to converge the input, or can be considered to be highly redundant across multiple noise realizations, as inherently uncertain. The use of this geometric uncertainty signal to estimate calibrated confidence is a new branch of research. Equally, in the large-scale multi task condition prevalent in foundation models fine-tuned to a wide range of downstream tasks, per task calibration differences have been found in even high quality shared representation setting indicating that

calibration is task sensitive in ways that are not well represented by the resultant confidence outputs of the shared model. All these findings imply that calibration study needs to develop out of its historical origins in image classification at a single task and continue to represent the totality of architectures, tasks and deployment conditions which comprise modern deep learning.

Summary Table 1: Comparative Overview of Calibration Methods

| Calibration Method | Category | Core Mechanism | Strengths | Limitations |
|---|---|---|---|---|
| Temperature Scaling | Post-hoc | Learns a single scalar T to rescale logits before softmax | Simple, fast, preserves accuracy; strong empirical performance | Single parameter may be insufficient for complex distributions |
| Platt Scaling | Post-hoc | Fits a logistic regression on model outputs using a held-out set | Well-established; effective for binary classifiers | Limited to binary or one-vs-rest settings; can overfit small datasets |
| Isotonic Regression | Post-hoc | Non-parametric monotone mapping fitted to predicted probabilities | Flexible, non-parametric; handles non-linear distortions | Requires large calibration sets; prone to overfitting on small data |
| Dirichlet Calibration | Post-hoc | Learns a full Dirichlet distribution mapping over class probabilities | Handles multiclass natively; captures inter-class relationships | Computationally heavier; more hyperparameters to tune |
| Label Smoothing | In-training | Softens one-hot targets during training to prevent overconfidence | Simple regularizer; reduces model overconfidence implicitly | Fixed smoothing factor may not match posterior distribution optimally |
| Mixup Augmentation | In-training | Trains on convex combinations of input–label pairs | Improves calibration and generalization simultaneously | Interpolation may produce out-of-distribution training samples |
| Monte Carlo Dropout | Bayesian-approx. | Keeps dropout active at inference; runs T forward passes | Provides uncertainty estimates without architectural changes | T forward passes increase inference cost; approximation quality varies |
| Deep Ensembles | Ensemble | Trains M independent networks; averages predictive distributions | State-of-the-art calibration and OOD detection performance | M× computational and memory cost; impractical for large models |
| SWAG (SWA-Gaussian) | Bayesian-approx. | Fits a Gaussian over SGD trajectory via | Better calibration than point estimates; single-model cost | Gaussian assumption may not capture |

| | | | | |
|---|---|---|---|---|
| | | stochastic weight averaging | | multimodal posteriors |
| Evidential Deep Learning | Direct estimation | Parameterizes a Dirichlet distribution as the output of the network | Distinguishes epistemic vs aleatoric uncertainty; efficient | Training instability reported; subjective choice of prior strength |
| Conformal Prediction | Distribution-free | Constructs coverage-guaranteed prediction sets using calibration scores | Rigorous finite-sample coverage guarantees; distribution-free | Produces sets, not calibrated scalars; efficiency depends on score function |

Summary Table 2: Uncertainty Quantification Frameworks for Deep Neural Networks

| **Framework / Method** | **Uncertainty Type** | **Inference Cost** | **OOD Detection Capability** | **Representative Works & Notes** |
|---|---|---|---|---|
| Deep Ensembles | Epistemic + Aleatoric | O(M×) | Excellent; consistently top-ranked on benchmarks | Lakshminarayanan et al. (2017); gold standard for UQ despite cost |
| MC Dropout | Epistemic (approx.) | O(T×) | Moderate; quality depends on dropout rate and architecture | Gal & Ghahramani (2016); widely adopted; underestimates uncertainty in some regimes |
| Variational Inference (BNN) | Epistemic | O(2×) per forward pass | Good if posterior approximation is accurate | Blundell et al. (2015); scalability remains a challenge for large networks |
| Stochastic Depth / DropConnect | Epistemic (approx.) | O(T×) | Moderate; similar limitations to MC Dropout | Effective for vision transformers; structurally implicit uncertainty |
| Natural-Gradient Variational Inference | Epistemic | O(1.5×–2×) | Good; tractable for convolutional and transformer architectures | Osawa et al. (2019); improved scalability over mean-field VI |
| SWAG / SWAG-Diagonal | Epistemic | O(1×) test | Good; captures flat minima curvature information | Maddox et al. (2019); practical for large models post-training |
| Laplace Approximation | Epistemic | O(1×) + Hessian | Good for in-distribution; weaker for far-OOD inputs | Daxberger et al. (2021); last-layer variant scales to large nets |
| Energy-Based OOD Detection | Epistemic (OOD) | O(1×) | Excellent; energy score separates in- and out-distribution well | Liu et al. (2020); simple score from log-sum-exp of logits |
| Conformal Prediction | Coverage-based | O(1×) + calibration | Good; conservative prediction sets grow for uncertain inputs | Angelopoulos & Bates (2022); rigorous marginal coverage guarantees |
| Evidential Neural Networks | Epistemic + Aleatoric | O(1×) | Moderate–Good; relies on prior strength and training stability | Sensoy et al. (2018); single forward pass for both uncertainty types |

| | | | | |
|---|---|---|---|---|
| Diffusion-based Uncertainty | Epistemic + Aleatoric | O(steps×) | Emerging; probabilistic generation enables natural UQ | Emerging trend in generative UQ; very high inference cost currently |

## 9. Open Challenges and Future Research Direction

Although it has made significant developments in the last ten years, model calibration and confidence estimation in neural networks are still research fields that lack any conclusion and have a number of unresolved problems. The initial and the most obvious is the lack of unanimously accepted evaluation metric: each of the various calibration measures comes with different methods, and the metric can invert the relative hierarchy of evaluation methods. An exercise to standardize calibration benchmarks, akin to the profiling that ImageNet offered to standardizing discriminative accuracy assessment, would have a dramatic impact and would enable reproducible comparisons to take place. Such benchmark would preferably consist of not only the static held-out test sets, but also dynamic evaluation protocols which test calibration under increasing distribution shift, selective prediction conditions as well as adversarial inputs.

The fact that uncertainty quantification methods can be scaled to very large models is a critical bottleneck. Although deep ensembles give state-of-the-art model calibration of models scaled to the ResNet-50 scale, further scaling of GPT-4-scale models (or Gemini-scale models) is both computationally and cost-prohibitive to most research groups. Among the most practically consequent research directions now underway is the development of parameter-efficient uncertainty quantification methods, including low-rank posterior approximations, uncertainty-aware adapters as well as prompt-based calibration, that can be applied to frozen or slightly fine-tuned foundation models, without retraining based on novel inputs. Principled optimization Theoretical gains in modeling the connection between model structure, training dynamics, and calibration behavior, at least in transformer-based models and their attention mechanisms, would give principled directions on how to design naturally more calibrated models.

The calibration relationship and other credible AI qualities, namely, fairness, robustness, interpretability, and privacy, should be more thoroughly and systematically explored theoretically and empirically. There is early evidence indicating that all these properties can be conflicting: ideas that enhance calibration can degrade resilience to adversarial examples, and equitable calibration can decrease the quality of global calibration. It is of significant theoretical and practical importance that a unified framework that views such desiderata as joint optimization objectives with clear demarcations of tradeoffs should be developed. Lastly, how to operationalize the concept of calibration to apply in the real world, such as how one can monitor calibration drift in a production system, how to instigate a recalibration when there is bad behavior, and how to convey the

uncertainty to human decision-makers in a manner understandable to the AI systems engineering community, is a relatively untapped area of bridging the gap between the technical research community and the AI systems engineering community.

## 10. Conclusion

The calculation of models and estimation of confidence is done not only by specialized statistical issues, but it is now a key part of the credible deep learning program. The chapter has followed the intellectual history of the field, beginning with the successive development of the foundations of probability theory and the experimental diagnosis of the miscalibration of modern architectures, on through the variety of post-hoc-based remediation approaches and systems using in-trainings, through the fertile terrain of quantifying uncertainty using Bayesian methods, conformal prediction, and the new issues arising with large language models and multimodal foundation systems. The main message arising is of productive complexity: calibration is not one problem with one answer, but a collection of similar properties probabilistic consistency, group-conditional reliability, coverage under distribution shift, epistemic-aleatoric decomposition, each one of which needs to be measured, to be measured to use different methodologies, to be evaluated using different protocols.

The most practical insights that can be distilled by the practitioners are as follows. The temperature scaling is a free and highly effective first line of defense to standard classification architecture, that is, it is computationally free. Deep ensembles give the best empirical guarantees in applications where the quantification of uncertainty actually needs a recalibration of confidence, whereas SWAG and Laplace approximations can be used at reduced cost. Conformal prediction is a theoretically grounded wrapping layer that is able to certify statistical reliability without making assumptions about the quality of the model. In large language models, calibration is an open and rapidly advancing research problem with the most promising active lines being verbalized uncertainty, consistency-based estimation and retrieval augmentation. As well as in any deployment concerning the safeguarded demographical organizations, in-group calibration audit is a pre-condition to dependable inference. Since deep learning systems continue to assume more consequential roles in the society, the development of uncertainty-aware, well-calibrated models is not only a technical aspiration but an ethical need.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 6: Robustness in Deep Learning: Noise, Perturbations, and Distribution Shifts

## 1. Introduction

Deep learning systems in safety-critical settings facing the real world have instructed robustness not as a fringe-level research question but as one of the motivations of the contemporary machine learning. Since neural networks are now integrated into the consequential decision pathways, including those of autonomous cars, clinical diagnosis, and financial fraud detection, as well as national security infrastructure, it is no longer negotiable that they will be reliable are given imperfect, adversarial or shifted conditions. However, there is an incessant and highly migrating discontinuity between the numerical demonstrations that are achieved in monitored laboratory standards as opposed to the compromised behaviours that the same alternatives have whenever such as the real world is engaged. This chapter logically reviews the theoretical and empirical environment of deep learning robustness, focusing specifically on the three related axes in which failures occur, namely, random noises and stochastic corruptions, structured adversarial perturbations, and distribution shift in the statistical distribution between training and inference conditions.

The challenge of robustness underlies well-learned concepts of deep neural networks, which are most interpreted in terms of the context of what deep neural learning. The seminal results published by Geirhos et al. (2019) showed that convolutional neural networks (CNNs) that have been trained on ImageNet generate a powerful texture bias, which is based on local statistics patterns, but can be highly fragile and susceptible to sensitive to surface textures manipulation, noise, compression artefacts, or style transfer. However, in comparison to human observers, such corruptions become tolerated with remarkable tolerance, using hierarchical, shape centric representations that carry across transformation. This discrepancy in the human and machine perception suggests that one of the defining characteristics of empirical risk minimization (ERM) is the fact that the optimization process fails to differentiate those features that are actually predictive of class identity and those that are simply correlated with labels in training distribution.

These spurious correlations break down when the test distribution is no longer at equilibrium, and errors can be dramatically, abruptly and unpredictably high.

To gain deep insight of the idea of robustness, it is important to draw the boundary of different sources of perturb stating that some of these sources are natural, as they are created by the data collection mechanism, whereas others are artificially caused with the purpose to take advantage of the weakness of the model. Natural corruptions, like blur, pixelation, weather artefacts, lighting variation, are a type of corruption considered of the former category, and investigated by Hendrycks and Dietterich (2019) as ImageNet-C and ImageNet-P benchmarks. Adversarial perturbations constitute a qualitatively distinct threat: initially designed and crafted by humans, imperceptible by the human eye, and deliberately selected to threaten a target model. Initially described by Szegedy et al. (2014) and then weaponized by Fast Gradient Sign Method (FGSM) of Goodfellow et al. (2015) and Projected Gradient Descent (PGD) attacks of Madry et al. (2018), adversarial examples can be regarded as one of the most examined and not completely resolved issues of trustworthy machine learning. The third axis Distribution shift refers to the statistical difference between the training and deployment data distributions and cuts across the phenomena as phenomena as diverse as covariate shift, label shift, and the smooth semantic drift that often typifies any temporal deployment situation. Children of these three axes determine the robustness landscape to which this chapter dedicates much specific mapping, the processes of robustness failure, and the increasing research in techniques by which robustness is also possible to be recovered or guaranteed.

Most recent developments, including large-scale pre-training and foundation models, have put a new spin on this debate. Models like CLIP (Contrastive Language-Image Pre-training), DINOv2, and large vision-language models have shown that training on a wide range of, web-scale corpora gives representations some amount of distributional invariance that was unattainable with other more supervised models, such as on curated collections. These observations have triggered some fundamental questions as to what a truly strong representation is and whether scale, diversity, self-supervised goals can be one of the paths to robustness leaving the constraints of adversarial training and explicit domain generalization methods. The chapter summarizes these recent views and old findings, outlines the situation in benchmark evaluation, and determines the gaps in the field that will constitute the frontier of robustness research by the year 2024-2025.

## 2. Theoretical Principles of Robustness

### 2.1 Formality and the Objective of Robustness

This is because formal definitions of the robustness of a deep learning model can be defined in terms of worst-case generalization. Since f: X - Y is a classifier that statistics X to Y takes input space X to label space Y and e > 0 is a perturbation budget measured in some norm * p. The adversarial robustness goal demands that, given any x in the distribution D, and all means of perturbing it, d, such that d p [?] e, the classifier is compatible with x in that it satisfies f(x + d) = f (x). This min-max model is the loss minimization over model parameters and the loss maximization over perturbations, and it is an instance of the adversarial training paradigm proposed by Madry et al. (2018) and offers a formal basis to certified robustness training approaches. The norm p adopted is the consequence of a norm p, the L[?] norm defining the maximum amount of per-pixel change, has been the most common choice in image classification studies because it corresponds to a more natural perceptual constraint that imperceptibly changeable perturbations should be pointwise small, whereas L2 and L0 norms are norms that measure the total energy and the sparsity of perturbation respectively, depending upon the nature of the threat considered.

In addition to the adversarial robustness, the more general statistical concept of the distributional robustness views the performance of the models based on their worst-case expected loss with respect to any distribution Q over an uncertainty set around the training distribution P. Distributionally Robust Optimization (DRO) as described by Duchi, Namkoong, and other researchers aim to compute parameters that contain the worst-case expected loss with respect to any distribution Q within the uncertainty set around the training distribution P. This framework deals with the covariate shift problem and the natural distribution shift problem which are discussed in latter sections. Theoretical characterization of DRO shows that there is an inherent conflict in that growing the uncertainty set makes the estimator more robust and guarantees that the estimator is accurate on average-case failed distribution, but also means that the estimator is more inaccurate on the original distribution, the nebulousness issue of the whole field. This trade-off, its reasons, and its being an inherent component or simply restriction of achievable training paradigms is a key open problem in the theory of robust machine learning.

### 2.2 The Trade-Off of Robustness-Accuracy

This is one of the most replicated findings in the robustness literature, the empirical observation that robustness to adversarial perturbations is generally associated with

poorer test-set performance. Tsipras et al. (2019) have described the theoretical background on this phenomenon and showed that in adversarial-trained models, learned features are more correlated with human-perceivable semantics, which is useful in interpretability and generalization, but did not necessarily achieve the best performance with clean data which follows the original distribution. It is intuited that adversarial robust models need to learn which have smooth decision boundaries that remain stable under small perturbation balls, and this smoothness limits the capacity of such models to draw on small-scale statistical dependencies that, although spurious, do fully predict these distributions (both in training and the standard test distribution). This was formalized by Zhang et al. (2019), who have introduced the TRADES (Tradeoff-inspired Adversarial DEfense via Surrogate-loss minimization) framework, making the explicit trade-off one of the design variables instead of an artifact of failure.

More recent neural evidence indicates that there is no upper limit to the extent to which this trade-off is true but that model capacity, pre-training quality, and data availability have a substantial influence on this assertion. Theoretically and empirically demonstrated by Raghunathan et al. (2020) is that unlabelled data can be employed to bridge the gap, which encourages a generation of semi-supervised and self-supervised methods in adversarial training. Vision models trained on billions of pairs of images and text have shown that at least when the pre-trained representation is rich enough the trade-off can be greatly softened indicating that data scale and data diversity may represent the strongest tool to jump up the robustness-accuracy Pareto frontier. The results have inspired the present research direction towards using foundation models as a strong backbone, and the task-specific fine-tuning to be used in a manner that does not harm the distributational invariance gained during pre-training.

## 3. Input Corruptions and Stochastic Noise

### 3.1 Taxonomy of Natural Corruptions

Natural input corruptions are the most pervasive cause of robustness challenges of deployed deep learning systems. These corruptions occur due to the physical constraints of the sensing devices, the varying of conditions of the environment as well as the corrupted nature of the data transmission and storage pipeline. The crucial benchmark characterization of this issue was given by Hendrycks and Dietterich (2019) ImageNet-C which is a collection of 75 different kinds of algorithmically generated corruptions at five severity indicators to the ImageNet validation set. These corruptions fall into four general categories namely noises corruptions (Gaussian, shot, impulse, Speckle), blur corruptions (Gaussian blur, defocus blur, motion blur, zoom blur), weather corruptions

(Fog, Frost, snow, brightness) and digital corruptions (JPEG compression, pixelation, elastic transformation). The canonical measure of natural robustness across the field has now become the mean Corruption Error (mCE commented as the average error of a model over all types of corruption divided by the error of a fixed AlexNet model).

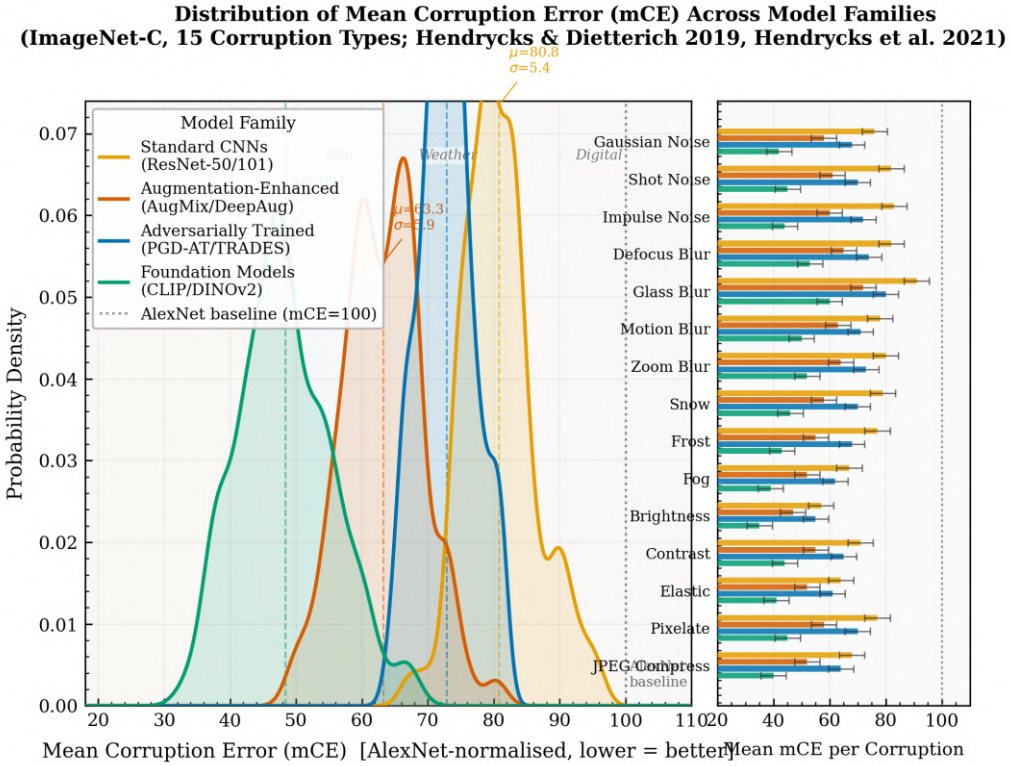


Fig.1 mCE KDE + Per-Corruption Bar Chart ·

In above fig.1 -Left panel: overlapping KDE density curves for four model families across all 15 ImageNet-C corruption types, with corruption super-category bands (Noise / Blur / Weather / Digital). Right panel: horizontal grouped bars with error bars showing mean mCE per corruption type per family.

There is no uniformity in the mechanisms through which the various types of noise undermine neural network performance and this heavily relies on the frequency properties of both the input signals as well as the learned filters. Gaussian noise, spreading energy evenly over the frequency energy, selectively hurts features trained on high-frequency channels, specifically those channels that most CNNs use as the ultimate providers of fine-grained discriminative information. Yin et al. (2019) showed that the Fourier analysis demonstrates that standard ResNet consumes high-frequency features which are effective to clean-data classification but are weak in real-world corruption, whereas data augmentation schemes to avoid high-frequency features illustrate low-

frequency and corruption-resistant representations. This Fourier view has been very influential in the future design of augmentation, leading to the frequency-domain approach to augmentation design and the application of anti-aliasing in network designs as a theory-guided approach to enhancing natural robustness.

### 3.2 Augmentation Strategy of Noise Robustness

Historically, data augmentation has served as the main defence against natural corruptions and recent years were marked by the explosion of sophisticated augmentation techniques that extend to the basic cases of geometric transformations. AugMix, which was proposed by Hendrycks et al. (2020) entails a combination of stochastic chains of elementary augmentation operations with a Jensen-Shannon consistency loss that motivates the model to make similar predictions to the image fragments with this loss altered in one way or another. This regularization of consistency is vital: augmentation-only trained models are more accurate on particular corruption categories but can increase corruptions on certain ones, and reduce corruptions on others, in consistency with the consistency goal does promote a more consistent representation which is more uniform. AugMix had significant gains on mCE on ImageNet-C with no decline or gain in clean accuracy, revealing itself to be a powerful baseline to natural robustness.

The DeepAugment model introduced by Hendrycks et al. (2021) as the benchmark augmentation model to the AugMax produces augmented training images by performing manipulations with the help of image-to-image neural networks, whereby the statistics of corruptions do not closely conform to a certain corruption type within the test set. This training method results in a more varied and randomized augmentation distribution that causes the model to learn purely invariant features as opposed to being overfitted to the noise distribution present at a particular point in training. In combination with these other augmentations and a stylized ImageNet training set, the given combination of augmentation cocktails achieves state-of-the-art natural robustness at the time of publication. The larger idea that to increase diversity and not aimed-at corruption is the ingredient that leads to natural robustness has been proven through several relevant studies and agrees with the distributional robustness approach which aims to achieve performance stability in the wide range of uncertainty, not through optimization against known corruptions.

A computationally efficient alternative is the RandAugment framework (because it selects uniformly at random an augmentation operation among a predefined library of operations), which has only a single hyperparameter the magnitude, and is controlled by a single hyperparameter (Cubuk et al., 2020). It is simple, and combined with competitive metrics in the classification, object detection and segmentation benchmarks,

has become the de facto standard augmentation strategy in large-scale training pipelines such as the EfficientNet training recipes and Vision Transformer (ViT) training recipes. Afterward, it was shown that the magnitude could to a tune of TrivialAugment (Muller and Hutter, 2021) be uniformly randomised, which also alleviates the augmentation design burden with further results showing the magnitude could be randomised as well. These processes represent a larger trend in favor of simple, scalable, and dataset-agnostic augmentation solutions as the practical base of natural robustness, with more elaborate approaches in cases where natural robustness is a major deployment condition.

## 4. Adverting Structural Perturbations and Structural Vulnerabilities

### 4.1 Adversarial example Phenomenon

Adversarial examples, which are inputs or constructions consisting of minor, carefully designed perturbations that ensure that a model makes incorrect predictions with high confidence, have essentially disrupted the assumptions around the trustworthiness of deep learning systems, as well as triggered a whole subfield devoted to such understanding and defense. It was first of all surprising when Szegedy et al. (2014) found that imperceptible perturbations could consistently deceive deep neural networks since observationally it appears that clean and adversarial images are visually similar. Goodfellow et al. (2015) later presented the hypothesis of linearization: the unnular dimensionality of the neural network activation in the input space implies that a tiny pertussis in the direction of the sign of the gradient of the loss conditioning (the FGSM direction) produces an atomic gradual modification in the output logits, adequate to switch the prediction of the design. Although the given explanation has been improved and to some degree replaced by geometric and information-theoretic views, the FGSM attack can be seen as a workhorse in the context of quick adversarial training and training robustness testing.

This is the Projected Gradient Descent (PGD) attack suggested by Madry et al. (2018) and is the golden standard at the time of assessing adversarial vulnerability in the context of L[?] threat model. PGD performs several steps of a gradient ascent to maximize the training loss on the e-ball centered around the input and projects the perturbed input back into the feasible set at the end of each step. Through repeated restarts of the process, PGD is a good and principled approximation to the worst-case perturbation an obstruction may reach with a first-order algorithm, and therefore a robust proxy of the actual worst-case adversary. Models trained on PGD attacks--also called PGD adversarial training--have the best empirical robustness of any known defence without formal certification guarantees, though, of course, at a significant cost in increased cost

of training (7-10x larger than normal training) and a non-trivial clean accuracy drop. The Madry et al. model of MNIST and CIFAR-10 provided a new model of defence evaluation in which reported robustness is shown to be provably resistant to adaptive adversarial attacks that are specially adapted to the defence mechanism.

### 4.2 The problem of Evaluation and Adaptive attacks

An evaluation problem in adversarial robustness research studies is that many of the proposed defences that seem robust to non-adaptive attacks are defeated by adaptive attacks that bypass the defence mechanism. This observation was recorded systematically by Carlini et al. (2019) through a review of a corpus of published defences and showing that most of them could be defeated by adaptive attacks with a small computational budget. This evaluation challenge was tackled by the AutoAttack benchmark (Croce and Hein, 2020) which built an ensemble of various attack algorithms, such as APGD-CE, APGD-DLR, FAB, and Square Attack, parameter free and thus offering a consistent and reproducible grading of adversarial robustness without the user having to manually adjust attack parameters. AutoAttack has since been used as the standard evaluation protocol of adversarial robustness on classification benchmarks, and a leaderboard RobustBench has been managed to maintain a constantly updated ranking of the most robust publicly available models on datasets and threat models.

Black-box adversarial attacks introduce a very different and highly practically relevant threat model where the adversary can not access the parameters or gradients of the model but can access the output of the model. Transfer-based attacks are based on the empirical observation that adversarial examples that are obtained against a given model often transfer to different models, especially where surrogate and target (or target-side) models belong to similar architecture families, or on similar training data. This transferability, which was actively investigated by Papernot et al. (2016) and later enhanced by intermediate-level perturbations (ILA) and momentum-boosted attacks, is a key systemic risk: in practice, adversaries in real deployements have no access to the parameters, but when transfer based-attacks are used to attack production systems, it can be done with only a locally trained substitute model. Black-box attacks based on decision and scores, such as Bondary Attack, ZOO, and RayS also extend the threat model by having more restrictive query only assumptions and are capable of an assaulting with thousands to hundreds of thousands model queries.

### 4.3 Certified Defences and Formal Guarantees

The failure of empirical defences to give guarantees against worst-case adversaries has due alone apparent motivations to the study of certified robustness-ways of giving

formal, mathematically provable bounds on the worst-case perturbation that a model can survive with to continue to make correct predictions. Cohen, Rosenfeld, and Kolter (2019), developed randomized smoothing, the most scalable certified robustness approach to large models: when the base classifier is f, a smoothed classifier g is trained by averaging predictions on Gaussian-perturbed examples of the input. When perturbing the L2 parameter it is possible to compute in the establishing of g the radius in which the prediction is approved to stay at equal footing, exactly in terms of the gap in probabilities at the two highest position classes of prediction. Randomized smoothing scales to ImageNet-class models and is the only known certification method that has ever been demonstrated on such scale, but suffers a certification cost in the form of the desired level of confidence and cannot be extended to L2 perturbations.

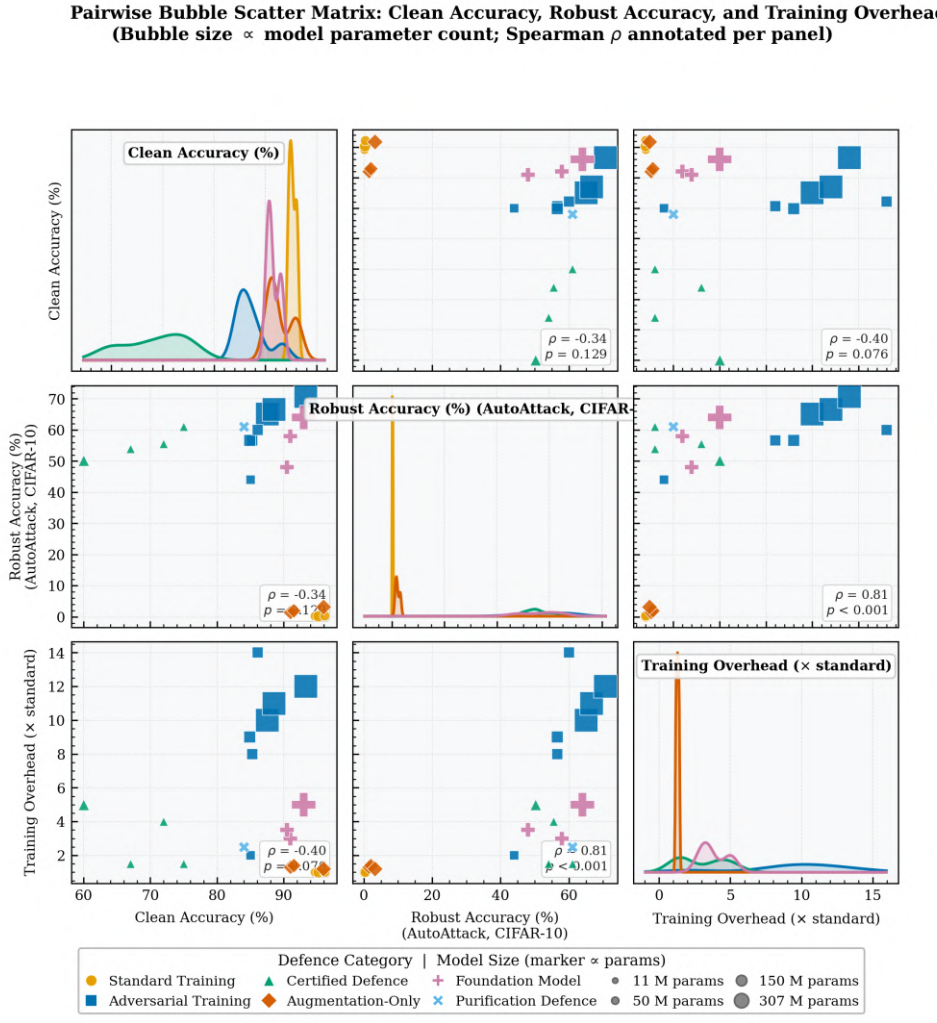


Fig.2 Pairwise Bubble Scatter Matrix

Fig. 2 is  A 3×3 matrix comparing Clean Accuracy, Robust Accuracy, and Training Overhead simultaneously. Diagonal panels show per-category marginal KDEs; off-diagonal panels show bubble scatters (size ∝ params) with Spearman ρ and *p*-value annotated in every panel.

Interval Bound Propagation (IBP) and its variations, such as the CROWN framework, a-CROWN, and b-CROWN, provide a deterministic way of certification by calculating tight bounds on activation intervals propagated across the network layers, and bounding the worst-case output logit gap of a given input perturbation set. These approaches have been demonstrated to have competitive certified accuracy on CIFAR-10 and MNIST, and generalized to transformers and graph neural networks by formulating narrower relaxations of attention and non-linear activation functions. The VNN-COMP competition has driven the fast development of neural network verification algorithms and a-b-CROWN has reached the state of the art in various verification challenges. The long-standing drawback of IBP-based approaches is that the loose intermediate bounds compound across the deep networks, and certified guarantee will be conservative and encourage recent studies to find stricter, and more computationally efficient relaxation approaches.

## 5.Shifts in Deep Learning Distribution

### 5.1 Types and Sources of Distribution Shift

Probably the most widespread of the robustness issues in the practice of machine learning is distribution shift, which represents the difference between the statistical characteristics of training data and those in the real-world setting at test or deployment time and occurs in virtually all applied fields. The conventional supervised learning paradigm follows that, test and training examples (xi, yi) follow independent, identical joint distribution, P(X, Y), the so-called i.i.d. assumption. The assumption is, in fact, violated in practically every deployment context where the training data has to be a finite historical sample of a distribution that has continued to change, which is geographically heterogeneous, and which is affected by selection biases that are inherent in data collection mechanisms. Distribution shift study aims at defining the kinds of distributional discrepancies which may arise, formulating conceptually sound indicators of the extent of the shift, and developing learning algorithms resistant to a particular type of shift expected in the application of interest.

The most widely researched subtype is called covariate shift, and indicates a change in the marginal distribution of inputs P(X) but whose conditional distribution is P(Y|X) is fixed. This is analogous to the intuitive case of where the underlying relationship

between inputs and labels is stable yet the categories of inputs at test time occur in a systematic manner when compared to the content of the training set say a medical imaging model trained on data in high-resource hospitals deployed in a facility with different scanner protocols and patient demographics. In covariate shift, the common empirical risk minimizer is not an inconsistent estimator of the target risk, and the asymptotic correction of importance weighting that corrects this inconsistency through reweighting the training examples by the ratio of the densities p_test(x)/p_train(x) is Asia theoretically consistent. Nevertheless, it is well known that it is challenging to use a statistical estimate with great accuracy in high-dimensional input spaces; and large density ratios cause high-variance estimators and thus can be hard to use to produce reliable estimates with a finite data set.

Label shift (or prior probability shift) here is a shift in the marginal distribution of the labels P (Y ) of the class-conditional distribution P(X|Y). Such problem occurs in medical screening tasks where the fraction of a certain condition after training is different between training time and trained model deployment, and where topic distributions and sentiment priors vary across time and platform with natural language processing. Lipton, Wang, and Smola (2018) introduced a computationally-efficient algorithm, the Black Box Shift Estimation (BBSE), algorithm, which can be used to estimate the shifted distribution of labels on unlabelled target data followed by correction of the predicted classifier results by changing the decision threshold accordingly. The most difficult type of shift is concept drift (i.e. the change of conditional P(Y|X) i.e. change in the true functional relationship between inputs and outputs over time is the most difficult type of shift to counteract, since it cannot be mitigated by reweighting, distributional calibration and usually needs retraining of models, or constant learning models.

### 5.2 Benchmarks for Distribution Shift

The creation of stringent measures of assessing the robustness of models when there is a shift in distribution has been one of the significant contributions of the robustness research community in the last five years. Koh et al. (2021) artificially selected 10 real-world distribution shift datasets, covering several different domains such as satellite imagery, medical histopathology, species classification, and sentiment analysis, with the offers of the WILDS benchmark. Most importantly, WILDS is based on naturally occurring distribution shifts, which are determined by subpopulation membership, geographic area, or time, as opposed to artificially-generated corruptions, which is much closer to the evaluation of robustness against the corruption that actually occurs in deployment. The benchmark is used to identify the difference between the in-distribution validation performance and the out-of-distribution test performance, and facilitate a controlled measure of the OOT generalization gap. WILDS results have continued to

show that standard ERM with data augmentation is a rather formidable baseline, even compared to many of the more advanced domain adaptation and domain generalization algorithms, a result that has dashed high hopes of complex shift-specific algorithms, and shifted the focus to more robust feature learning algorithms.

The problem of ImageNet distribution shift has mushroomed with the recognition that ImageNet validation accuracy is not a reliable measure of robustness in the real world. ImageNet-V2 (Recht et al., 2019) gathered a fresh test set in accordance with the initial ImageNet collection approach except observed that all designs show a uniform accuracy decrease of 11-13 percentage points of the original test set distribution was not visible through typical performance. ImageNet-Sketch, ImageNet-R (Renditions) and ObjectNet test models are tested with more extreme domain failures, namely, photographic style, artistic rendition, and controlled viewpoint and background variation respectively. ImageNet-A (Natural Adversarial Examples) is a collection of natural images in the real world that can cause these models to fail frequently with no artificial adversarial example input, which is a natural complement to algorithmically generated adversarial examples. The relative change in the aggregate of these benchmarks is uniform: existing deep learning dominate models are highly fragile to numerous natural distribution changes and ImageNet validation performance correlates poorly with robustness validation performance.

### 5.3 Domain Independent Learning and Generalization.

The domain generalization algorithms attempt to use multiple source domains to train models that become useful to novel target domains without any target domain information. Invariant Risk Minimization (IRM, Arjovsky et al., 2020) postulates the existence of robustly predictive features, whose optimal empirical classifier is uniform across training conditions formally as the gradient of the environment-specific loss of empirical classifiers with respect to a common representation should equal zero at a common representation. The point of the matter here is that, although spurious correlations can be useful predictors in the context of particular training environments, they will not pass this invariance test in a variety of different environments, and thus can be filtered out. Though IRM and its variants (IRMv1, IRM Games, VREx) have been proven theoretically valid, empirical conclusions on real-world problems have been inconsistent, and proceeding analysis by Rosenfeld et al. (2021) established that linear IRM is susceptible to failure under certain circumstances. The empirical comparison of domain generalization algorithms proposed by DomainBed benchmark (Gulrajani and Lopez-Paz, 2021), has verified that the domain generalization advantage of specialized algorithms over a finely tuned ERM baseline has consistently been small and unstable

across datasets, making it difficult to reliably solve the problem of domain generalization.

Test-time adaptation (TTA) techniques adopt an alternative approach and only use unlabelled data of the target domain by adapting model parameters/statistics during inference. TENT (Wang et al., 2021) optimizes the batch normalization layer statistics and affine parameters to minimize the entropy of the predictions on test batches and the fast and efficient adaptation is achievable without any source data access. Later applications have applied TTA to continuous and on-line (CoTTA, ETA), to full parameter adaptation with close regularization, and to the more difficult single-sample adaptation (TTT++) case where no test batch exists. One problem with TTA methods is that there is error condensation and representation breakdown in case of large or rapid distribution shifts, which is an incentive to the memory-efficient and collapse-preventing mechanisms. Dynamic Uncertainty Adaptation (DUA) technique and MEMO have solved the single-instance adaptation through creation of augmented views of every test sample followed by maximization of marginal entropy which serves as a viable point between a batch-level TTA and the single-instance adaptation needed in most real-world tasks.

## 6. New Methods and Current Innovations.

### 6.1 Diffusion Models as Defences

The diffusion-based generative models have led to the development of a new adversarial defence research front as these models are incredibly successful. DiffPure suggests the forward-reverse diffusion process (Nie et al., 2022), consisting of pre-processing, where the adversarial perturbation is removed by diffusing the adversarial input forward through time, and the adversarial structure is diminished by the addition of Gaussian noise, and then unravelling is performed by reversing diffusion to reconstruct a no-adversarial distribution image that the classifier is applied to. This method delivers state-of-the-art performance on conventional adversarial benchmarks on some threats of interest and has the theoretical advantage of not altering the classifier itself, therefore being applicable to any deployed model as a plug-and-play defence. The difficulty of adaptive evaluation is especially severe in the case of diffusion-based defences: some adversaries can use alternatives to the diffusion process as differentiable surrogate objectives, which means that they can produce more powerful attacks that pass the purification stage. More recent efforts in certified purification, and in tighter adaptive attacks, have made clear what the conditions are that DiffPure can give any genuine robustness guarantees other than to obscure vulnerabilities due to gradient masking.

### 6.2 Foundation Models and emergent Robustness.

Recent breakthroughs in the ability to train robust base models on multimodal data on web scale have redefined the robustness space in previously unanticipated ways that the classical paradigm of robustness studies does not predict. Trained on pairs of images and texts using contrasting samples of 400 million examples of internet images, C fortified sharp transfers on the distribution shift benchmarks, becoming significantly, much more effective on image classification under a range of distribution shifts than supervised models trained directly on ImageNet (where despite it never being fine-tuned). The analysis by Wortsman et al. (2022) later indicated that linear probing or fine-tuning CLIP features and interpolates (weight-space ensembling) between the fine-tuned and zero-shot model parameters has better accuracy-robustness Pareto performance than either extreme, indicating that the pre-trained features are distributional invariant, which are partially destroyed by fine-tuning. The results obtained by DINOv2 (Oquab et al., 2023), and optimized self-supervised vision transformer trained on a curated dataset of 142 million images (distillation goal) support the hypothesis that the quality of the representations and the diversity of the training data are the leading factors in determining the robustness: the model achieves competitive out-of-distribution performance on a number of different benchmarks without any task-specific supervision.

The complex robustness profile in large language models (LLMs) and multimodal large language models (MLLM) is different. Although GPT-4, Claude, and others are quite impressive when it comes to the range of natural language tasks, they are susceptible to adversarial prompt injection, jailbreaking, and semantic perturbations that semantically fix the surface form, despite creating a change in form. The sensitivity of LLMs to prompt formatting, phrasing of instructions, and selection of examples in context further complicates the problem of robustness as such images cannot be evaluated in the same way as robustness in image classification, and instead need new assessment procedures. Evaluations like AdvGLUE, Prompt Bench and the adversarial evaluation suites created by the MLCommons AI Safety working group have only recently been used as benchmarks to evaluate the robustness of LLMs, where the rate of model development always outpaced the creation of detailed evaluation protocols. In multimodal systems, the application of perturbations to the visual stream can robustly control the textual instructions, which is unsafe because systems are used in agentic functions where visual and textual messages are both used to decide on high-stakes actions.

### 6.3. Effective Self-Supervised and Contrastive Learning

Self-supervised (SSL) algorithms in which the objective of the learning process is based on contrastive or non-contrastive terms have proven to be competitive in natural robustness relative to supervised algorithms with the benefit of scalability through the

avoidance of human-supervised labels. An especially notable result is that the SS models that learn with high data augmentation, which is the augmentation-invariance property that drives the use of contrastive goals, intuitively learn representation that is resistant to the nature of the corruptions applied to the data during augmentation. SimCLR, MoCo, BYOL, and Barlow Twins All have been demonstrated to perform as well as or better than supervised representations based on ImageNet-C and ObjectNet with similar architecture and training complexity.

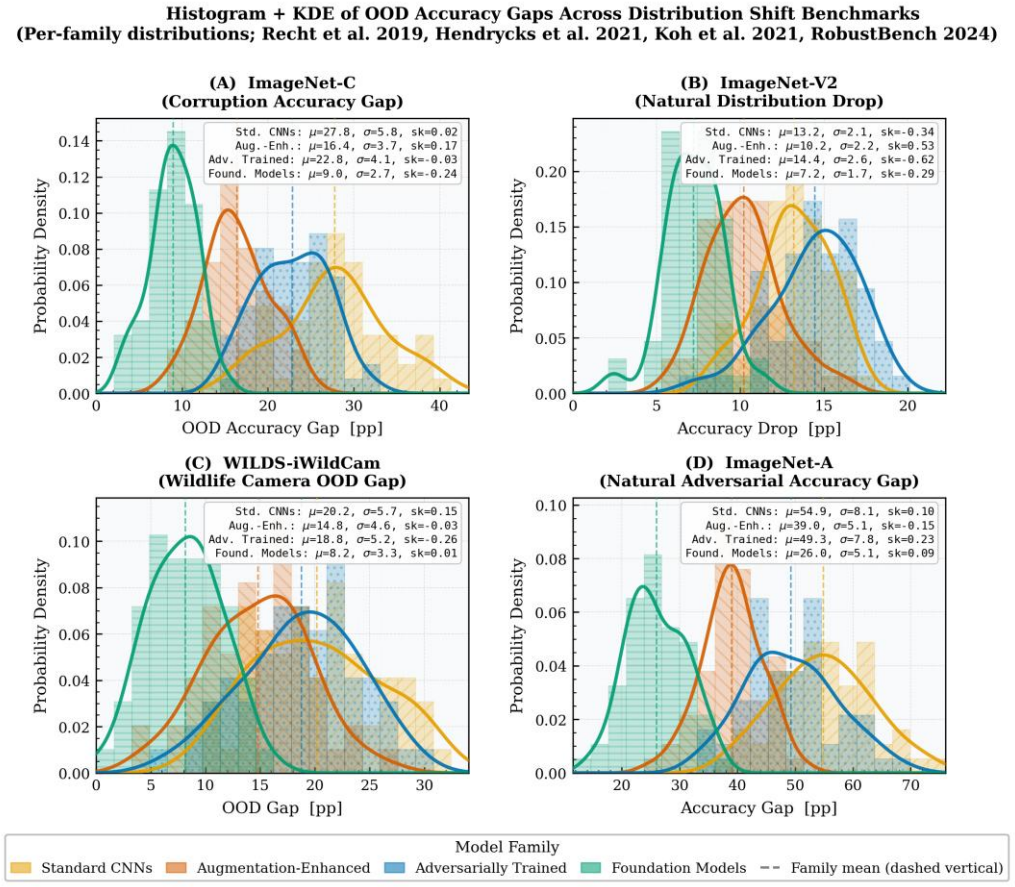


Fig.3 OOD Accuracy Gap: 4-Panel Histogram + KDE

In above Fig.3 One panel per benchmark (ImageNet-C, ImageNet-V2, WILDS-iWildCam, ImageNet-A), each showing hatched histograms overlaid with KDE curves for four model families, plus a statistics box (μ, σ, skewness) per family.

This has inspired studies on adversarially robust contrastive learning where adversarial perturbations are also used as an extreme example of data augmentation and consistency between clean and adversarial views is also enforced as a pre-training goal. The study of adversarial robustness on representations has shown that adversarial robustness can be

self-supervised, and downstream adversarial robustness is enhanced by several factors relative to other models, making it possible that the pre-training phase is an effective and under-used technique to gain robustness in the representation, as opposed to just fixing such representations along the fine-tuning phase.

Table 2.1: Taxonomy of Robustness Challenges in Deep Learning — Sources, Affected Models, Metrics, and Mitigations

| Challenge Type | Source / Origin | Representative Models Affected | Key Metrics | Mitigation Approaches |
|---|---|---|---|---|
| **Gaussian Noise** | Sensor imperfections, environmental stochasticity | ResNet, VGG, EfficientNet | Corruption Error (CE), mCE | Augmentation, noise injection, spectral normalization |
| **Impulse / Salt-and-Pepper Noise** | Communication channel errors, sensor faults | CNNs, Transformers (ViT) | Peak SNR degradation, mCE | Median filtering, robust loss functions |
| **Adversarial Perturbations (white-box)** | Gradient-based attacks (PGD, FGSM) | All gradient-based architectures | Adversarial accuracy, Attack Success Rate | Adversarial training, certified defences, randomized smoothing |
| **Adversarial Perturbations (black-box)** | Transfer-based, decision-boundary attacks | Deployed APIs, black-box models | Transferability rate, query efficiency | Input transformation, ensemble defences |
| **Natural Distribution Shift** | Domain gap, geographic/demographic variance | ResNet, BERT, GPT variants | ImageNet-C accuracy, OOD gap | Domain generalization, invariant risk minimization |
| **Covariate Shift** | Change in input feature distribution P(X) | Tabular ML, medical imaging models | KL divergence, Maximum Mean Discrepancy | Importance weighting, batch normalization variants |
| **Label Shift / Prior Probability Shift** | Change in class prior P(Y) | Classification heads, NLP classifiers | ELCE, calibration ECE | Label shift correction, BBSE, RLLS algorithms |
| **Semantic Shift (Concept Drift)** | Evolving task definition over time | Sequential decision models, LLMs | Drift detection latency, accuracy decay rate | Continual learning, test-time adaptation, DUA |
| **Texture / Style Perturbations** | Artistic style transfer, rendering engine gaps | CNNs (texture-biased), ViTs | Shape-texture bias index | Stylized ImageNet training, data augmentation (AugMix) |
| **Geometric / Spatial Transformations** | Camera angle, rotation, scale changes | Object detectors, pose estimators | Spatial robustness score, mPC | Geometric augmentation, spatial |

## 7. Performance Review Structures, Measures, and Standards.

Robustness evaluation in deep learning has developed beyond the customary reporting of accuracy on concrete assaults to an advanced vicinity of standardized benchmarks, examination actions, and automated leaderboards that allow to assess the model robustness across various threat models and situation of distribution shift in reliable ways. Setting up evaluation criteria has been extremely significant in the achievement of scientific rigor in an area in which the antagonistic character of the problem has provided a robust motivation towards benchmark-driven optimization and cherry-picked evaluation procedures. RobustBench, which is being upheld by Croce et al. and where Auto Attack forms an evaluation protocol and to which both standard accuracy and robust accuracy are reported on standard image classification benchmarks under both the threat models L[∞] and L2 is now the canonical reference point in adversarial robustness research. The best publicly available models on CIFAR-10 at the L[∞] constraint with ε = 8/255 currently have a robust accuracy of about 71 percent and a clean accuracy of about 95 percent, a gap that measures the price of robustness currently and the way to improve it.

In addition to adversarial robustness, distributional robustness needs measures that can represent the various causes of distribution shift that happen in practice. The five-fold design of the WILDS benchmark, in which in/out-of-distribution validation, in/out-of-distribution test, in/out-of-distribution test, and a different test in another domain are distinguished, gives a more subtle evaluation framework allowing to differentiate a model with strong in-distribution performance and a model with uniformly strong performance across worst-case subgroups. The primary measure in WILDS when equity of performance across demographic subgroups is a deployment requirement is rather worst-group accuracy than average accuracy. This is a measurable set of objectives of fairness indicating that the understanding of robustness and fairness are increasingly seen as tightly linked: in models with good average performance and disastrous performance on underrepresented subgroups, is not robust or fair, and solutions to the subgroup robustness problem, like Group DRO and Just Train Twice, tend to combine methods of both the robustness and fairness literatures.

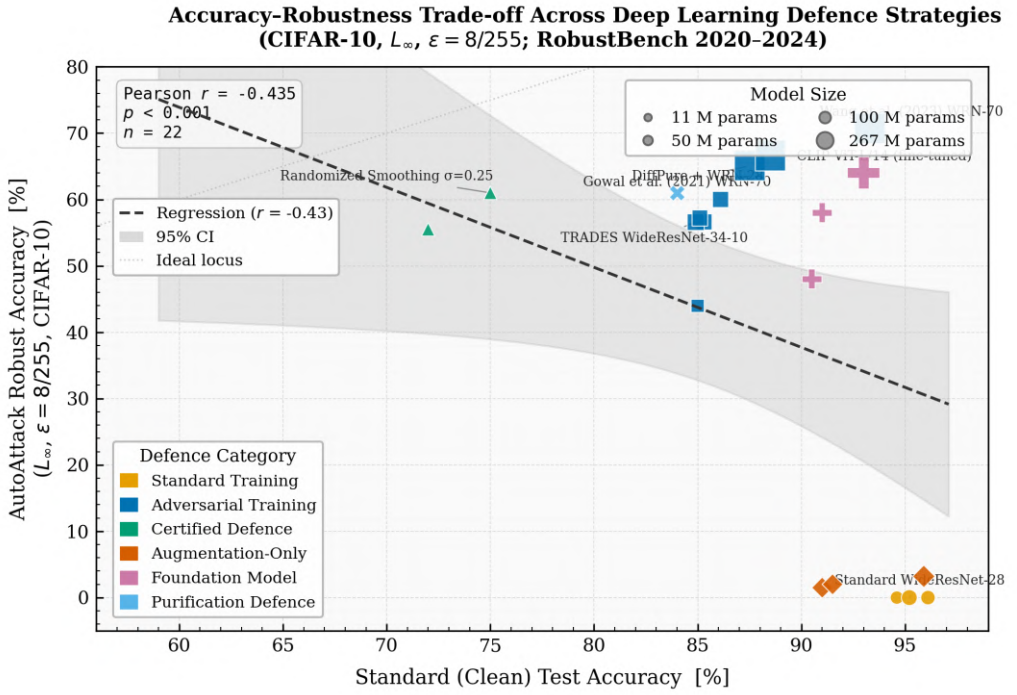


Fig 4. Accuracy–Robustness Trade-off Scatter

Above Fig.4 describes clean accuracy vs. AutoAttack robust accuracy for 22 models on CIFAR-10 (L∞, ε=8/255). Marker shape = defence category, size ∝ parameter count. Linear regression with 95% CI envelope and Pearson r annotation expose the canonical negative trade-off.

The concept of robustness with respect to large language models poses methodological special problems that current benchmark infrastructure is not well-prepared to cope with. Output space of the generative models is continuous and unlimited and binary accuracy metrics cannot be applied. Classes of semantic equivalence one of the sets of outputs that a human judge would consider equally valid or acceptable are very hard to do formally and are costly to mark. The problem of automated evaluation with the LLM judges is that they are circular with the possibility of distribution shift in the judge model. These issues have pressed the need to invest in human evaluation research, adversarial red-teaming, as well as automated evaluation systems that compound semantic similarity measures with both fact-checking and activation safety constraints. The HELM (Holistic Evaluation of Language Models) benchmark proposed a multi-metric multi scenario evaluation system that evaluations in terms of accuracy, calibration, robustness, fairness, efficiency, and safety and offers a more detailed portrait of model trustworthiness than unidimensional benchmarks.

## 8. Future Reimbursement and Open Problems.

Deep learning robustness represents a productive, unsolved frontier with notable empirical advances in particular settings as well as theoretical foundational problems that have been open after approximately ten years of active research efforts. There are various new trends that will transform the scenario significantly in the next few years. One, introducing robustness goals as an integral part of pre-training large foundation models as opposed to considering the task of robustness as a downstream defence task: The introduction of robustness goals as a central task of pre-training large foundation models is a paradigm shift with potentially transformative outcomes. More recent theoretical investigations have positively indicated that self-supervised goals of contrastive and non-contrastive loss, and can be assumed to have implicit distributional robustness, and assuming that this property can be enhanced by the judicious design of goals and through data curation, it may be possible to build foundation models whose representations are strongly invariant to changes in the natural distribution when deployed, without necessarily which adversarial objectives have been trained. The scaling assumption of robustness, which larger models which are trained on more diverse data are systematically more robust, needs strict empirical exploration and theoretical foundation.

Second, the strengths of contemporary agentic and thinking systems in terms of interacting with tools, with external APIs and with real-world situations present new challenges that far surpass the those of image classification threat models that have dominated robustness research. In agentic systems, adversarial lesions can be added to the retrieved files, web pages, or tool results, and the results of one prediction error can propagate through multi-step inferences chain in qualitatively new ways than a single error of the classification. Immediacy walks, adversarial instrumentation, and situational control over chain-of-thought rationale represent novel threat vectors that need solidness investigation to obtain into the extent of the true complexity of real-world LLM execution. On-the-film verification of LLM outputs, specification-based robustness check, agentic-specific red-teaming protocols are under development and the quality, when these reach maturity, will be key to safe deployment of more autonomous AI systems.

Third, the connection between robustness and efficient inference is an empirical issue that is increasingly becoming a challenge. Most of the most effective robustness techniques such as adversarial training, randomized smoothing, diffusion purification, and ensemble defences have very high computational overhead which is hard to balance with the inference latency of edge deployment, real-time systems, and resource-constrained systems. Priority research directions including the development of efficient certified robustness techniques, hardware-aware adversarial training, and test-time adaptation algorithms with sub-linear complexity that will decide the useful operation of

robustness guarantees in deployed systems deem to be practical. Neural architecture search based on neural architecture search with robustness objectives and the creation of intrinsically robust architectural primitives, including group-equivariant convolutions, capsule networks and hyperdimensional computing, provide other means of achieving robustness without the computational cost of training-time defence algorithms, by encoding robustness as an inductive bias.

Lastly, the discipline should take the evaluation gap--the continued difference between robustness score on curated benchmarks and the actual robustness in deploying the system into practice very seriously. Benchmark overfitting, i.e. the tuning of robustness procedures in implicit or explicit designs to particular sets of threat models and types of corruption, is a true threat that has already become apparent in the natural robustness literature where ImageNet-C-specific augmentation strategies are found to be unreliable under other switching benchmarks. Differentiation of evaluation protocols adversarial to evaluation - benchmarks that are constructed to be challenging to overfit, benchmarks that contain corruption modes one is not aimed at, and benchmarks that are regularly challenged - is a significant meta-research agenda. The general movement towards more concrete deployment testing, human testing and domain specific robustness testing will eventually be required to bridge the divide between robustness studies and practical trustworthiness requirements such as those of actual AI systems.

Table 2.2: Robustness Enhancement Methods — Principles, Performance, and Limitations

| Method / Technique | Category | Underlying Principle | Empirical Performance Gains | Limitations & Open Challenges |
|---|---|---|---|---|
| **Adversarial Training (PGD-AT)** | Training-time defence | Min-max optimization over worst-case perturbations | ~50–65% robust accuracy on CIFAR-10 (epsilon=8/255) | High training cost; accuracy–robustness trade-off |
| **Randomized Smoothing** | Certified robustness | Monte Carlo certification via Gaussian smoothing | Certifiable L2 radii; scalable to ImageNet | Soft accuracy penalty; inapplicable to large perturbations |
| **AugMix** | Data augmentation | Stochastic augmentation chains + consistency loss | 25–30% mCE reduction on ImageNet-C | Limited against adversarial perturbations |
| **DomainBed / Invariant Risk Minimization (IRM)** | Domain generalization | Learns invariant features across training environments | Consistent OOD gains on PACS, OfficeHome | Sensitive to environment labelling quality |
| **Test-Time Adaptation (TENT / DUA)** | Inference-time adaptation | Entropy minimization on batch normalization statistics | 3–8% accuracy recovery under shift | Unstable under large shifts; requires mini-batches |

| | | | | |
|---|---|---|---|---|
| **Certified Defences via Interval Bound Propagation (IBP)** | Formal verification | Propagate perturbation bounds through network layers | Verifiable guarantees; scales to medium networks | Conservative bounds; accuracy degradation |
| **Denoising Diffusion Purification (DiffPure)** | Pre-processing defence | Forward–reverse diffusion to remove adversarial noise | State-of-art under adaptive attacks (2023–2024) | Inference latency; diffusion model dependency |
| **Foundation Model Fine-Tuning (CLIP, DINOv2)** | Pre-training robustness | Large-scale diverse pre-training captures robust features | Significant zero-shot OOD generalization gains | Computational cost; potential spurious correlations |
| **Vision-Language Robustness (LAION training)** | Multimodal robustness | Align visual and textual embeddings for invariant representations | Improved texture/style robustness across benchmarks | Sensitivity to prompt engineering; hallucinations |
| **Ensemble Adversarial Training** | Training-time defence | Augment training data with adversarial examples from multiple models | Improved black-box robustness vs. single-model AT | Ensemble overhead; coordinated attack vulnerability |

## 9. Conclusion

This chapter has presented an in-depth investigation of robustness in deep learning in the three basic axes that fragility can be observed, namely stochastic input noises and natural corruptions, structured adversarial noise, and statistical distribution shift. We have followed through the theoretical underpinnings of each area of problem, the formulated min-max robustness objective and its relation to the distributional robustness optimization, the empirical landscape of adversarial attacks and certified defences, to the domain generalization and domain test-time methods of mitigating these problems that have been devised to cope with them. The main challenge types and responses to the methods are condensed in the two summary tables, which form a systematic guide to practitioners going through the literature on robustness.

There are a number of cross-cutting themes that come out as a result of this analysis. The robustness-accuracy trade-off is an empirical yet flexible constraint defining limited capacity of the models, data variability, and training paradigm, and more recent results of large scale foundation models indicate that the Pareto frontier of this trade-off can be pushed to the limits many times when pre-training conditions are rich enough. The issue of evaluation--that robustness assertions are not a product of an unfinished or a non-adaptive testing mix--is as critical as the defining issue of defence and it needs to be invested in further through the development of standardized adversarial-to-the-evaluator

benchmarks. The emergent class of agentic, multimodal, and reasoning systems presents qualitatively new robustness challenges, to which the available toolkit does not provide the full capability to respond to, and to which the available red-teaming technique must be used to characterize.

The risk of a failure in robustness also grows as the deep learning systems take on more responsibilities and autonomy in their work. An X-ray machine that malfunctions when subjected to sensor noise, a health care system that malfunctions when presented with a patient belonging to an underrepresented population, or an autonomous agent that can be influenced with adversarially engineered inputs is not merely the X-ray machine that simply runs on incorrect outputs, it can even be harmful to people themselves and undermine the trust in the society that is the condition to the effective implementation of AI. The broad sense of robustness as introduced in this chapter is thus not only a technical attribute that the developer has to optimize in a race condition but in itself is a fundamental aspect of social contract between AI systems and the citizens they support. In the following chapters of this book, complementary views of the quantification of uncertainty and adversarial resilience will be formed along with this chapter giving a single basis on the application of trustworthy deep learning.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 7: Adversarial Machine Learning: Attacks and Defense Mechanisms

## Abstract

Adversarial machine learning has become one of the most impactful subfields of current artificial intelligence research, which lies at the border of security, statistics, and deep learning theory. An in-depth analytically rigorous study of the adversarial threat picture that faces modern deep neural networks features the use of the newest theoretical results and empirical methodologies published until 2024 and early 2025 to accomplish this purpose. The chapter starts by locating adversarial vulnerability in more general terms as the generalisation behaviour of high-dimensional, over-parameterised models trained on finite data and why high-dimensional models are easily vulnerable to imperceptible, yet reliable, input perturbations that induce catastrophic misclassification. It takes a disciplined taxonomy of category of attacks, beginning with the classic white-box gradient-based algorithms like the Fast Gradient Sign Method and Projected Gradient Descent and continuing on to the latest decision based, physical world and large language model prompt-injection attacks, before switching to a correspondingly systematic scan over defensive strategies. Some of the studies considered are adversarial training, certified robustness methods, based on randomised smoothing and interval bound propagation, diffusion-model-based adversarial purification, architectural solutions, and game-theoretic forms of the attacker-defender relationship. This chapter ends by integrating the discussion of the empirical robustness-accuracy trade-off, open theoretical issues, and the practical implications of adversarial machine learning in high-stakes deployment settings of healthcare, autonomous systems, financial technology and state-scale language modelling. The main types of attack and defence mechanisms are summarized in two summary tables to allow a scholar to find this information swiftly.

## 1.Introduction

The impressive performance of deep neural networks in vision, language, audio, and decision-making algorithms has led them to be the cornerstones of the contemporary applications of the artificial intelligence system. Since diagnostic radiology systems that compete with specialist clinicians to huge language models that can reason in multi-steps, these structures have demonstrated their ability to leak intricate statistical regulators out of large amount of data. But it is precisely this success that has attracted newly scientific attention to an all too fragile sort of virtue: it can now be generated with astonishing ease and with remarkable calculability and determinism to arrange input perturbations small enough that these can cause confident misclassification- or arbitrary behavioural manipulation- of state-of-the-art neural models. Formally described by Szegedy et al. (2014), in the case of image classifiers, the phenomenon in question, later proven in practically all forms of deep learning modality, is known as the domain of adversarial machine learning.

The concept of adversarial machine learning is not a niche theoretical academic interest. Since neural models are setting up shop in research and development labs or are becoming central components of autonomous vehicles, medical diagnosis pipelines, financial fraud detection engines, or content moderation platforms and military decision-support systems, the fact that they contain reliably exploitable adversarial vulnerabilities is deeply worrying on the question of safety, fairness, and societal trust. Even a single adversarial patch on the stop sign can make a vehicle-mounted object detector label it as a speed-limits sign; any imperceptible modification to the image of a chest X-ray can mislabel an AI diagnostic system as having given a cancer diagnosis, not a malignant image; a promotes a language model can jailbreak a large language model, instead giving safety-consistent outputs; any text-based user prompt can jailbreak a large language model, with any operating point of the language model. Such cases are not hypothetical constructs, all of them have been proved in peer-reviewed journals, and some of them have been implemented into working systems.

The chapter is written in a way to target not only the advanced student who is being introduced to adversarial machine learning but also the established researcher who needs to have a review of what the state of the art really is on the subject matter. Section 2 includes the theoretical background required, one of which is the formal definition of adversarial examples, the threat models on which one classifies attacks, and the geometric intuitions that can be drawn on the loss landscape of neural networks. Section 3 provides a detailed taxonomy of adversarial attacks, including white-box gradient-based, black-box and transfer-based, physical-world perturbation, backdoor and data poisoning, and also adversarial attacks on natural language processing, methods. Section 4 also provides a survey of the defence scenarios in the same level of depth including empirical and certified defences, input preprocessing, adversarial purification,

architectural robustness, and game-theoretic formulations. The summary tables are illustrated in section 5. The cross-cutting themes mentioned in Section 6 encompass are the robustness-accuracy trade-off, the transferability phenomenon, certified radius limitations, and implications on trustworthy deployment. The conclusion of (7) is open research directions.

## 2.Theoretical Underpinnings of Adversarial Vulnerability.

### 2.1 Advocacy of Formal Defined Adversarial Examples.

An adversarial example is a formal definition of an adversarial example, i.e. an input x' that is formed based on a valid input x with the properties (i) a distance measure d(x, x') falls below a given perturbation budget e of some norm, typically the l [?], l2 or l0 norm, and (ii) a classifier f(*) predicts x' to fall into a target class y. The perturbation budget e is an attempt to represent an imperceptibility level: perturbations that do not exceed the norm of perturbation l[?] in the domain of pixel intensities [0,1] are customary to ImageNet benchmarks and represent the level of perturbation that can be perceived by human viewers under standard viewing conditions. Although the l[?] norm has been the most important actor in the literature on adversarial robustness, more recent work has studied the extension of the threat concept to semantic norms, a perceptual similarity metric like LPIPS (Learned Perceptual Image Patch Similarity), and the Wasserstein distance between the image distribution which have recognised that geometrical norm-balls are only partially explanatory of human perceptual invariance.

The statistical learning theory approach to the existence of adversarial examples, on the one hand, can be explained by the phenomenon of concentration of measure in high-dimensional spaces. In a d-dimensional hypercube the volume of any of the thin shells on the surface is a disastrously large proportion of the total volume as d gets large. It can be seen that these neural network decision boundaries, required to divide this high-dimensional space to represent thousands of categories, are geometrical nigh to most natural data-points not due to the poor training of the classifier, but as a consequence of the geometry of high-dimensional spaces. This intuition was formalized by Gillmer et al. who showed that a perfectly Bayes-optimal classifier set on a natural distribution will nevertheless be vulnerable to adversarial examples in case the set of data has a high intrinsic dimension. This finding would be very discouraging on theoretical grounds: it implies that the issue of adversarial vulnerability goes beyond bad training or undercapacity, and may be an unalterable characteristic of high-dimensional learning.

## 2.2 Systems, Threat and Attack Taxonomy Requirement.

Characterisation of adversarial attacks rigorously needs the threat model, the formal assumptions regarding the adversarial information and capability. Knowledge of a threat model Attacker knowledge A threat model that is black-box is more susceptible to attack than one that is white-box. Inference-time versus training-time A threat model that is norm-bounded is more vulnerable to attack than one that is norm-unbound. The white-box model is known to be an adversary having full access to the architecture, parameters, and gradients of the model, which is a stronger condition than any other that maximises the effectiveness of attacks and is suitable in the context of worst-case robustness. The black-box model only accepts queries by the adversary, producing either soft probability score or hard predicted labels, and thus gradient based optimisation cannot take place, requiring other methods like a transfer based attack, evolutionary algorithm or finite difference estimation of gradients. The practical significance of the difference between score-based and decision-based black-box attacks is that it is particularly in the interest of many commercial machine learning APIs to discourage query-based machine learning attacks by explicitly minimizing score values when scoring: hard-label attack algorithms, including HopSkipJump and the Boundary Attack, are developed.

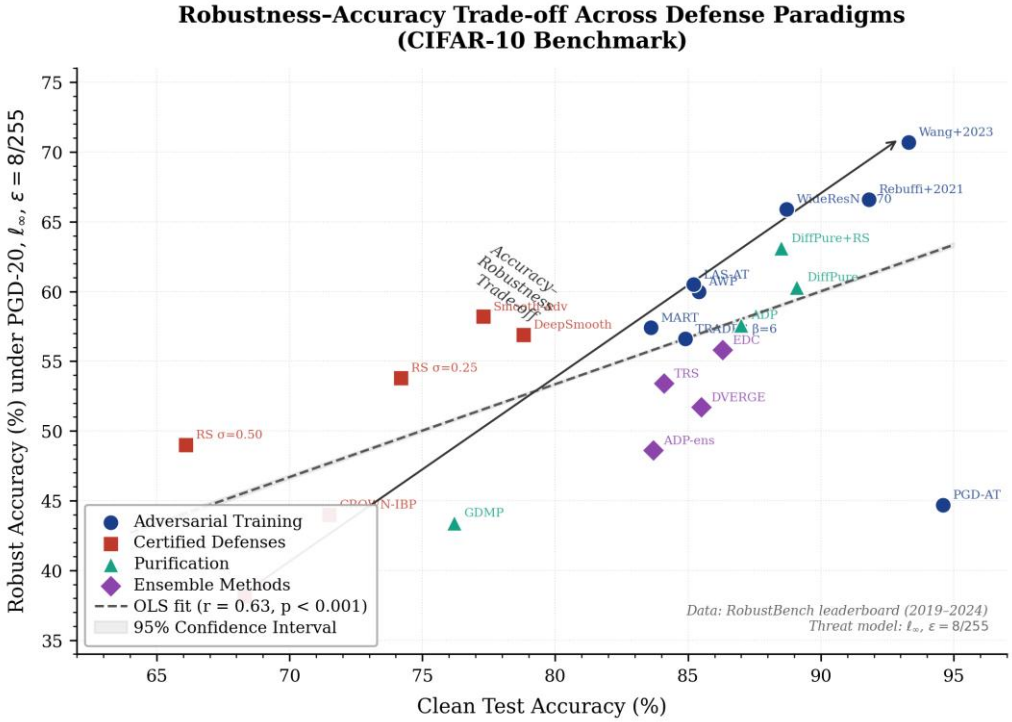


Fig.1 Robustness–Accuracy Trade-off Scatter Plot

Fig.1 shows A pairwise scatter plot of clean test accuracy (x-axis) vs. robust accuracy (y-axis) for 22 defense methods on CIFAR-10 (PGD-20, $\ell\infty$, $\varepsilon$=8/255). Four defense categories—adversarial training, certified defenses, purification, and ensembles—use distinct markers and colors. An OLS regression line with a 95% confidence interval band is overlaid, yielding the correlation coefficient and a trade-off arrow that visually enforces the accuracy–robustness tension documented in Tsipras et al. (2019).

The adversarial training model based on an entirely different and significantly more threatening threat model, since such adversarial training interventions continue to exist in the parameters of the model and not in a single inference. Backdoor or Trojan attacks Backdoor attacks and Trojan attacks consist of corrupting training data so that the resulting model will still act normally on censored data, but behave adversely whenever it is presented with a particular triggering pattern the appearance of a small patch, a colour filter or even a semantic attribute like the presence of sun glasses. The backdoor threat model includes compromising of the supply-chain, in which a practitioner trains a pre-trained model to act in a particular manner, acquired via a mistrusted repository, and federated learning, in which a malicious actor introduces poisoned gradient changes that get incorporated into a widely shared model. The two scenarios represent realistic attack surfaces of modern machine learning deployment pipelines.

## 3.Adversarial Attack Methodologies.

### 3.1 Gradient-Based Attacks on the White Box.

A white-box adversarial attack Featherstone et al. (2019) presented a method used as the foundational white-box adversarial attack: Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014): the method builds an adversarial example in a single step by altering co-determined input dimension e with the gradient gap of the classification loss to the input: The sign of the gradient of the classification loss (L) with respect to the input. Although FGSM was conceptually simple, it demonstrated that the linear nature of state-of-the-art deep networks with respect to high-dimensional input spaces, which makes them useful in training networks based on gradients, is also a severe security drawback. Kurakin et al. generalised FGSM to iterative variants, which apply a small step size a and project back onto the e-ball with each step, resulting in the Basic Iterative Method (BIM) followed by the Projected Gradient Descent (PGD) attack by Madry et al. (2018) with random initialisation included. Adversarial examples generated by the PGD are now considered the standard test of adversarial robustness: training a classifier on worst-case PGD loss, a method also called PGD adversarial training, is one of the

most effective empirical defenses against l[?]-bound perturbations, but comes at a significant cost in terms of training and also exhibits a non-trivial clean accuracy loss.

The Carlini and Wagner (C&W) attack reformulates the adversarial example generation problem as a constrained optimisation problem wherein the objective functional is a classification loss function with an added term of the magnitude of the perturbation (where the adversarial perturbation must cross the decision boundary) to ensure the adversary operates under the constraint of the smallness of the perturbation. Changing the variables: The C&W attack works in the logit space, parameterising the perturbed by the hyperbolic tangent function and optimises with Adam, generating adversarial examples which have significantly smaller l2 distortions as compared to FGSM or PGD with the same misclassification confidence. At the time of publication, the C&W attack was able to beat most defences suggested at time, and the then popular defensive distillation technique, effectively providing a measure of the challenge it poses to an adversarial defence problem. DeepFool, suggested by Moosavi-Dezfogh et al., is a geometric algorithm, which repeatedly projects the input onto the closest classification boundary, which is approximated by linearising the classifier, resulting in near-minimal perturbations to discover the local geometry of the loss surface.

### 3.2 Black-Box, transfer and query based attacks.

Transferability is the empirical finding, first to be systematically reported by Szegedy et al. and since then a subject of extensive further research, that adversarial examples constructed against one network may lead to misclassification when applied to a second network trained on the same task, although the specifics of the architecture, initialisation or hyperparameters of the two networks may be quite different. This fact is a concern as far as security is concerned since it implies that a malicious user who has access to a surrogate model can nevertheless pose a threat to a black-box target model by designing transferable adversarial examples over the surrogate. The recent momentum iterative fast gradient sign method (MI-FGSM) of Dong et al. improved the transfer rates significantly by introducing a momentum term which anchors the optimisation trajectory, and makes the adversarial perturbation unduly sharp along directions particular to the surrogate model. Further improvements to the DI-FGSM and TI-FGSM family were introduced that employed random input transformations at every iteration and convolve the gradient with a translation-invariant kernel respectively, both aimed at reducing the surrogate-specific fingerprint of the adversarial perturbation.

Black-box attack Query based attacks remove the transferability assumption and rely on directly estimating the gradient of the loss of the target model by applying well-informed queries. Zeroth-Order Optimisation (ZOO) attack generates an approximation of gradients with finite differences by querying the model at points generated by a

difference between pairs of points in each input dimension, which is a computationally costly approach to high-resolution images, and depends on the number of queries scaled with the dimensionality of the input. Natural Evolution Strategies (NES)-based attacks are a more fundamental probabilistic model of gradient estimation, which models the direction of search an expectation with respect to a perturbation distribution and approximates it with the use of antithetic sampling. The square attack by Andriushchenko et al. (2020), total non-gradient-based, rather than garden-blindly searching gradient perturbations, but in the square shape, by taking advantage of the perturbation space structure as a means to attain competitive attack success rate with significantly fewer queries than gradient-estimation algorithms. Decision-based attacks, whose operation assumes a black-box environment in which no more information than the predicted class label is known, investigate the geometrical boundary of the classifier by performing a random walk, starting at a starting point that it can be confidently misclassified, and back to the original input at a minimum of distortion. HopSkipJump and Sign-OPT are current state of the art in this regime having near-minimal distortions, few queries and are a combination of binary searching in the decision-boundary and gradient sign estimation.

### 3.3 Universally and Physically-Adversarial Perturbations.

The whole paradigm of physical-world adversarial attacks, introduced by Kurakin et al. and widely generalized by the work thereafter, calls the belief in the existence of adversarial perturbations as solely digital manipulations of pixels in a fixed resolution representation into question. Adversaries can also then use adversarial textures by printing them onto real objects, affix adversarial patches to surfaces, or by wearing adversarial designed eyeglass to delude classifiers trained on goods taken by a real camera at different times, under different lighting, viewpoints and distances. Eykholt et al. proved that when moving adversarial stickers to physical stop signs, object detectors always falsely classified them despite changes in viewing angle and distance in ways that would have washed out digital perturbations. This is generalised to a completely localised attack in the adversarial patch paradigm introduced by Brown et al. they can a single printed patch somewhere on an image--which need not be on the object under attack--grouped together by the convolutional neural network to spatial position and use it anywhere to dominate the prediction. Adversarial audio examples demonstrated by Carlini et al. have the potential to order voice assistants and be heard like music or normal speech by humans, and the implication of this is clear on the security of smart speakers, as the physical presence of an adversarial threat is now evident in the audio space.

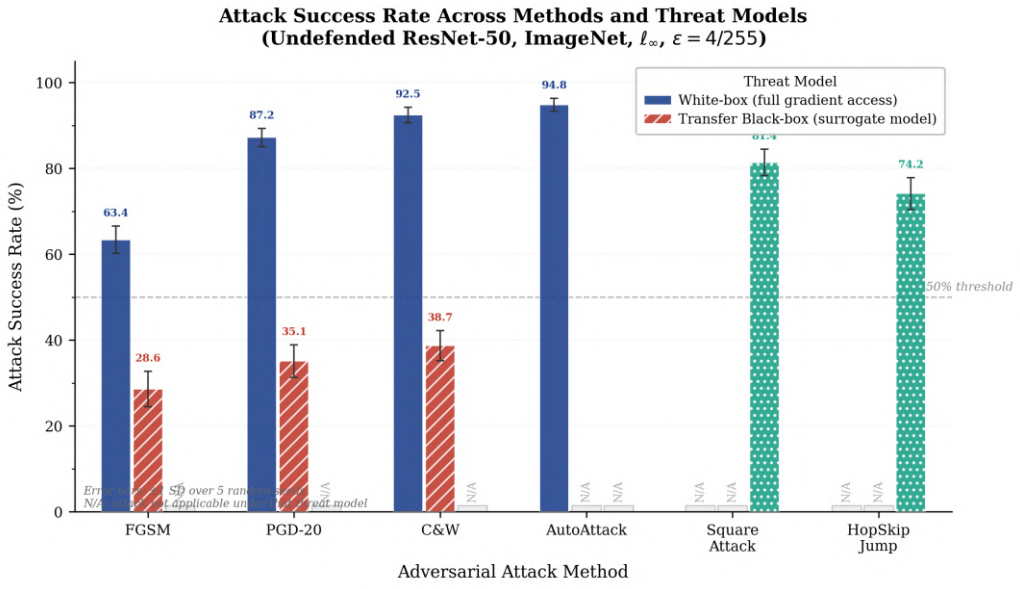


Fig.2 Attack Success Rate Grouped Bar Chart

Above fig.2 is A grouped bar chart with ±1 SD error bars comparing the attack success rate of six canonical adversarial methods (FGSM → HopSkipJump) across three threat models: white-box, transfer black-box, and query black-box. Inapplicable combinations are displayed as shaded "N/A" stubs with explicit labeling, while a 50% dashed reference line demarcates practically effective attacks from weaker ones.

Universal Adversarial Perturbations (UAPs), proposed by Moosavi-Dezfogh et al. (2017) are also an example of image-agnostic perturbation: a single fixed perturbation vector, applied to almost every image in the dataset, will trigger a high likelihood of misclassification. The fact that UAPs exist is conceptually astounding, - it means that the set of natural images, with all the enormous diversity they possess, have an overlying common directional weakness in the input space of this model and that this directional weakness is systematically exploitable. Later research has found that UAPs can be calculated without information about the training data (data-free UAPs) and can be produced in such a way that they cause certain classes to be misclassified as opposed to being generically falsely misclassified, and that they transfer very well among models. Practical risk of UAPs is extensive: in a surveillance system with the live video stream being processed, any uniformly applied single adversarial frame to the video stream will stop the person detector (person detection needs no per-frame computation).

### 3.4 Backdoor Attacks and Data Poisoning

Training-time adversarial manipulation Backdoor attacks, also known as Trojan attacks, is a type of backdoor attack, a training-time attack that commits the adversary to insert a latent narrative into the model: the model behaves normally in clean input conditions, but behaves according to the instructions of the adversary agent in conditions in which the adversary triggers a backdoor. This was experimentally shown in the seminal BadNets attack (Gu et al., 2017) where a simple pixel-pattern trigger was used in image classifiers; it was subsequently dramatically extended in both the sophistication and stealthiness of trigger, to both visible patches to invisible feature-space triggers (ISSBA, FTrojan), semantic triggers such as bald hair, sunglasses, and frequency-domain triggers that cannot be seen spatially. TrojanNN revealed that Trojan backdoors can be added to a model after it has been trained by specific manipulation of weights and no alterations would have to be made to the training data, which is an especially worrisome instance of attacker behavior to the pre-trained models acquired at open repositories used in the supply-chain. Bagdasaryan et al. also showed model-replacement attacks when a malicious party replaces its local gradient update with an adversarially-efficient instance that overrides the global model behaviour on trigger inputs and mixes into the overall update where Byzantine-robust aggregation conditions involving median or trimmed mean are inadequate in the federated learning setting.

### 3.5 Specific Natural Language Processing and Large Language Model Adversarial Attacks.

This threat to adversarial machine learning has grown far beyond computer vision systems to include natural language processing systems, such as the large language models which currently form the basis of much deployed AI products. Adversarial examples Adversarial examples in NLP systems are advantageously defined by a unique feature of having the input space discrete, i.e. text is made up of tokens in a finite vocabulary, preventing the use of continuous optimisation methods in one step in the image domain (leading to adversarial examples). Jin et al. (2020) and Li et al. (2020) can avoid this by greedily generated adversarial sentences based on semantic similarity constraints and context-sensitive embedding distances but generate grammatically correct and semantically coherent adversarial sentences that effectively and reliably mislead BERT-based classifiers but are nevertheless correctly classified by human readers. Character and subword Character-level and subword-level attacks take advantage of the artefacts of tokenisation of transformer models, and place homoglyphs, diacritical marks, or invisible Unicode characters, which do not have an impact on human readability, and alters model predictions.

The development of generative large language models has cannonized the prompt injection attack as a new and distinct adversarial threat vector that is a deadly no pun

intended. An immediate injection attack loads adversarial instructions in the inputs accepted by an LLM-based system in a manner that makes the model give precedence to the injected instructions over those of the legitimate instructions accepted by the system, with the consequence of leaking sensitive work or generating damaging material, or altering the tool-use or agent behaviour. It was demonstrated by the Greedy Coordinate Gradient (GCG) attack of Zou et al (2023) that adversarial suffixes calculated with gradient-based optimisation with open-weight models will transfer with worrying success to closed, safety-aligned commercial models, effectively jailbreaking these systems and avoiding safety training. These results have led to intensive work into the strength of reinforcement learning based on human feedback (RLHF) alignment processes, recent results indicate that adversarial alignment circumvention attacks identify gaps in the existing paradigm of preference-based safety training.

## 4. Defence against Adversarial attacks.

### 4.1 Adversarial Training and Its Recent Variants

The most empirically successful type of defence against adversarial attack is adversarial training, and is the basis of practically robust deep learning. This idea consists in augmenting the training goal with examples that are adversarially perturbed at test time, making the model mitigate the loss at the worst-case of examples perturbed by the threat model instead of minimising the average-case loss on clean examples. This is formalised by the PGD adversarial training formulation of Madry et al. (2018), and presented as a min-max optimiser: the inner maximisation compiles the strongest adversarial example in the e-ball at the current training step, by use of multiple iterations of VPGD, and the outer problem is to modify the parameters of the model best able to classify this worst case example. Although this is effective, vanilla PGD training has various pathologies which have been well documented: it loses clean accuracy significantly (the so-called clean-robust accuracy trade-off), and is computationally costly because of the inner optimisation loop, and tends to be vulnerable to l2 or semantic perturbations.

TRADES (Zhang et al., 2019) is a principled reformulation of adversarial training, in which the robust error is reformulated into a natural error term and a penalty of the sensitivity of the models prediction on the perturbed data inside the e-ball as a KL divergence term. Through this decomposition is the explicit robustness-accuracy trade-off is revealed under the control of a hyperparameter b: as b increases, particular focus is given to robustness with the price of clean accuracy, and vice versa. TRADES always exhibits higher performance on the clean robust Pareto frontier compared to PGD adversarial training on a variety of benchmark and architectures. MART (Wang et al.,

2020) is also an improvement over TRADES in three ways: it differentiates between false and true classified adversarial examples in the loss, uses a larger regularisation on adversarial examples already near the boundary, and empirically shows that such bias of adversarial examples on the boundary results in better robustness. Adversarial Weight Perturbation (AWP, Wu et al., 2020) further includes a dual perturbation strategy which additionally perturbs the model weights in addition to the input and is aware that flat loss landscapes on model parameters indicate larger robustness margins and that the act of perturbing weights during training represents an implicit version of sharpness-conscious regularisation. The state-of-the-art adversarially trained models on CIFAR-10, with a robust accuracy range of 67-70% against PGD-e=8/255 attacks are, as usual, very expensive to achieve, and even with this level of robustness are now a significant distance from the 94-96% clean accuracy of the corresponding unperturbed models, highlighting the high cost of empirical robustness.

### 4.2 The Certified Defences and Formal Verification

The empirical defences that are effective against the known attacks do not imply the effectiveness against the adaptive adversaries that are capable of maximising their attacks targeting the particular mechanism of defence. The adversarial machine learning literature has reported the phenomenon of broken defences on multiple occasions: The defences that seemed to be good against standard attacks were then proved to be circumvented by adaptive attacks that are explicitly optimised against the defensive. It is an epistemological challenge that drives certified defences, which give mathematical proofs that any perturbation to the model within a given norm ball cannot alter the prediction of the model, based on the adversary's strategy. The most practically scalable certified deep neural network defense, which is currently the subject of study by Cohen et al. (2019), is randomised smoothing, introduced by Lecuyer et al. (2019) and refined extensively by Cohen et al. (2019): which consists of building a smoothed classifier g(x) as a plurality vote of a base classifier f, defined at copies of x noised with Gaussian noise and a variance s, and proves via the Neyman-Pearson lemma that g is certifiably robust in Certified to ImageNet resolution and arbitrary network Architectures, randomised smoothing scores ImageNet, with certified radii at l2 scale, as competitive with empirical defences on CIFAR-10, nevertheless, the inherent noise injection causes significant clean performance loss at the noise levels needed to achieve significantly certified radii.

Interval Bound Propagation (IBP) as well as linear relaxation-based methods: CROWN, DeepPoly, and a-CROWN, supply certified robustness bounds a value obtained by running the network on symbolic representations of the perturbation set, calculating an upper bound on the worst-case loss in the perturbation ball. CROWN and its variants make use of the linearity of ReLU networks (piece-wise) to develop tight linear bounds

on the activation value of each neuron, and it produces much tighter certificates compared to naive intervals arithmetic. Combined with training using certified training objectives, these approaches produce provably robust classifiers, which have certification overheads that are orders of magnitude less than full verification using mixed-integer linear programming, but are only scalable to small models and low input dimensionality in comparison to the size of production systems. An alternative to certified robustness, provided by the Lipschitz network paradigm, is to tighten spectral normalisation, gradient penalty regularisation, or new network architectures that allow tight Lipschitz bounds, including Cayley-parameterised orthogonal networks. An Lipschitz constant L network is certified sturdy to perturbations of size e in the l2 norm in the occasion the spreadiest between the highest-ranking classifier and the subsequent best ranker is greater than 2Le; but maintaining small Lipschitz constants and roughly keeping the classification error on complicated data reflects a fundamentally difficult architectural difficulty.

### 4.3 Preprocessing of Input and Adversarial Purifying

Preprocessing defences (based on input preprocessing) seek to eliminate or mitigate adversarial noise at the inputs before it arrives at the classifier, having the benefit of relying on the structural difference between fully optimised adversarial noise (maximally offensive to the classifier) and natural image statistics. Xu et al. (2018) postulate that such reductions will cause adversarial perturbations designed in the full-resolution, full-bit-depth input space to be significantly mitigated by attempts to reduce the colour bit depth of the input in feature squeezing, which postulates that the adversarial noise was designed to be minimised by the reductions through the use of spatial-smoothing filters. JPEG compression also eliminates high-frequency adversarial components also making use of the perceptual quality measure that JPEG quantisation is optimised. Although these techniques make such common attacks less effective, more general adaptive attacks, in which the preprocessing step is explicitly specified as part of the adversarial optimisation problem, as a differentiable approximation can usually find adversarial examples that pass through the preprocessing and greatly impair the efficacy of the method as a defence on its own. Encoding by thermometers, defences based on randomisation, and total variation denoising have also been suggested and it has been demonstrated that they too can be defeated by suitably adaptive adversaries.

One of the most recent hits in the defence scenario is adversarial purification that is based on deep generative models, especially denoising diffusion probabilistic models (DDPMs). DiffPure (Nie et al., 2022) offers to clarify the adversarial examples by diffusing them forward through the DDPM Markov chain in a limited number of steps (adding controlled Gaussian noise) and reversing the diffusion process (denoising) by

taking advantage of the fact that the reverse diffusion process that a well-trained score functions guide the noisy adversarial input back onto the natural data manifold. The diffusion process is a stochastic process, which adds noise at each step of diffusion, and thus forms a non-differentiating barrier that thwarts gradient-based adaptive attacks, which empirically demonstrates a significantly larger improvement over deterministic preprocessing approaches. Adjoint Method A more precise gradient estimation facilitated by the diffusion ODE solver is introduced by Zhang et al. (2023), and later on showed that purification and randomised smoothing can both be used to empirically and certifiedly robust at the same time. The key weaknesses of diffusion-based purification include the very large inference-time overhead, that is, executing the diffusion denoising model takes dozens to hundreds of network forward pass, and the fact that the purified image can contain semantically significant modifications to the image, which is changing the ground-truth label.

### 4.4 Architectural Strength and Characteristics Learning Strategies

An entirely different attempt at adversarial resistance is that of constructing neural networks exhibiting a suitable geometry and low adversarial resistance as an architectural attribute, not as a post-training method or during inference. The fact that vision transformers (ViTs) are more adversarial robust than convolutional neural networks in various studies was originally due to the fact that multi-head self-attention layers operate on fixed patches of the input image and does not use local texture statistics but instead uses global contextual features. This observation has since been confirmed by systematic analysis, sometimes with a significant degree of qualification: ViTs are not as susceptible to texture-based adversarial attacks, but much more susceptible to patch-based, adversarial colour jitter, frequency-domain attacks, indicating that not all architectural biases perform well on adversarial vulnerability. CoAtNet and CvT are hybrid convolutional-attention models that show to be able to combine the local inductive bias of the convolutions with the global contextual processing of the attention and which seem to be more robust to a greater variety of types of attacks.

Strong feature learning methods tackle adversarial vulnerability on the representation level, which aims at following feature extractors that learn semantically robust characteristics on the input perceptors, but not some surface level statistical artefact. The feature denoising networks combine non-local mean denoising blocks in the layers of feature extraction, and the adversarial noise elements are weakened in network representations, eliminated at the bottom representations. Inspired by causal inference theory, Invariant Risk Minimisation (IRM) and its variants learn a representation that is invariant under changes of environment or information generating distribution, and

conjecture that features predictive in distribution shift are less sensitive to adversarial manipulation since adversarial examples are a certain manifestation of distribution shifts. Contrastive adversarial training is a variant of adversarial training along with self-supervised contrastive learning goals, where the network is trained with the InfoNCE loss to entice the representation of an adversarial example to be geometrical to the representation of its clean counterpart of the embedding space, which is essentially adversarial invariant feature extraction. Such techniques show a promising future especially given the fact that by pre-training on large unlabelled datasets followed by adversarial fine-tuning has been demonstrated to lead to significantly better robust generalisation.

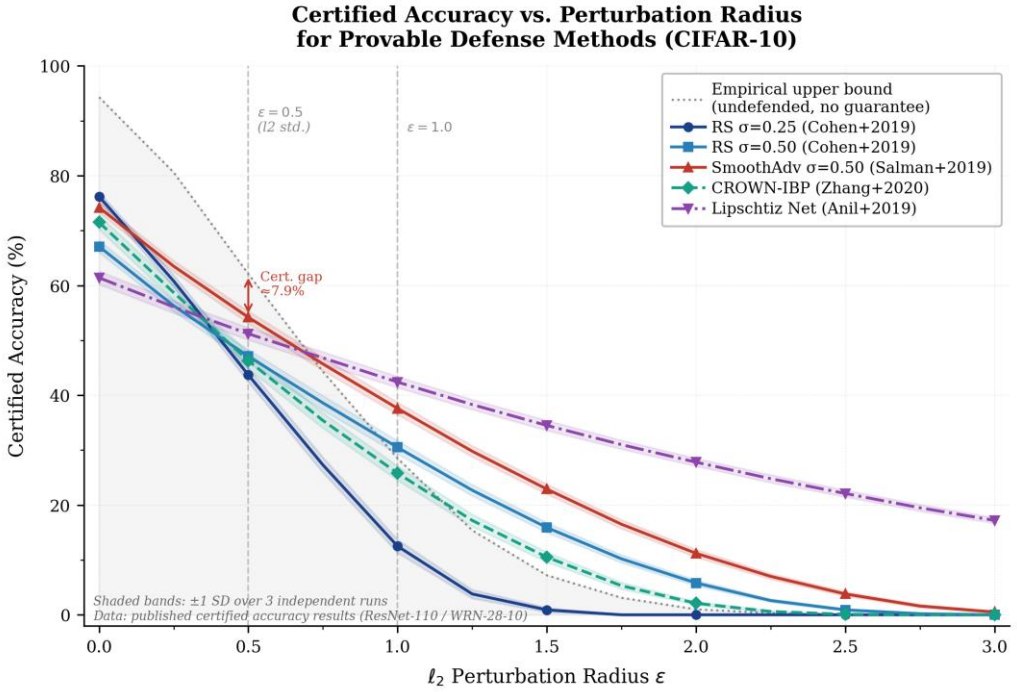


Fig.3 Certified Accuracy vs. Perturbation Radius

Fig.3 describes A multi-line plot with shaded ±1 SD confidence bands showing how certified accuracy degrades as the $\ell_2$ perturbation radius ε grows, for five provable defense methods on CIFAR-10. The empirical undefended upper bound is plotted as a reference, and a bidirectional annotation arrow marks the certification gap at the standard ε=0.5 benchmark — the difference between what is empirically achievable and what can be formally proved.

### 4.5 Detection Game-Theoretic Formulations, Ensemble Methods.

Instead of making individual classifier more robust, detection-based defences introduce a detector of adversarial examples that tries to determine whether the input to a system is adversarial or not, before it can be sent to a classifier, allowing the system to handle the fallback case, seek clarification, or trigger a fallback mechanism. Local Intrinsic Dimensionality (LID) detection, Mahalanobis distance-based detection and deep k-nearest neighbour detection all have the guessed property when adversarial examples are identified, they will occupy distinct regions in the model interior feature space they will tend to cluster around the boundary of the classes, they will tend to exhibit an abnormally high local intrinsic dimensionality, and they will not lie within the support of the natural training distribution as measured by the intermediate layer statistics. Obviously, these detectors exhibit high scores on the AUROC score when benchmarked on standard attacks, but these detectors are equally vulnerable to detection-aware adaptive attacks, which use the detection mechanism as a part of the adversarial objective and choosing offensive perturbations to cause misclassification and evade detection simultaneously. Ensemble defences which rely on Adaptive Diversity Promoting (ADP) regularisation and DVERGE (Decorrelated Vulnerability for Diverse Ensemble) aim to make an ensemble of models the adversarial vulnerability of an ensemble is decorrelated, allowing an adversarial instance to nullify an individual model, yet other instances counteract the nullification of other adversaries of the ensemble: the ensemble vote becomes hard even when any individual vote is not. DVERGE applies vulnerability decorrelation by minimising distillation-based loss which reduces the transferability of adversarial example to members of an ensemble.

The game theory descriptions of adversarial machine learning present the engagement of an adversary and a defender as a two participant zero-sum Stackelberg game in which the latter reveals the initial move first to a training algorithm and the former subsequently responds optimally. Such an approach will give a clean rationale as to why adversarial training can be seen as seeking a Nash equilibrium of the min-max game, and why oblivious defences, i.e. those whose design is independent of adaptive adversaries, will always be beaten by adaptive adversaries who will best-respond. More recent studies have generalized this model to multi-round and multi-attacker designs, where adversarial vulnerabilities are considered through the deployment lifecycle of a model, as opposed to an interaction between adversaries and defenders in a single round. Distributionally robust optimisation (DRO) offers an alternative theoretical grounding, where the classifier is trained to minimise the worst-case expected loss on a mixture of training distribution adversarially picked, which is a formulation that generalizes adversarial training and domain adaptation to special cases. Adversarial settings that have been optimized with group DRO and Wasserstein DRO have been used to obtain more

robustness guarantees that go beyond the norm-ball pertussis model but are coming with the computational complexity cost of a barrier to large-scale application.

## 5. Summary Tables

Table 1. Taxonomy of Principal Adversarial Attack Categories: Methods, Threat Models, Exploited Vulnerabilities, and Application Domains

| Attack Category | Representative Methods | Threat Model | Key Vulnerability Exploited | Primary Application Domain |
|---|---|---|---|---|
| White-Box Gradient Attacks | FGSM, PGD, C&W, DeepFool | Full model access (weights & gradients) | Gradient sign direction of loss surface | Image classification, object detection |
| Black-Box Transfer Attacks | MI-FGSM, DI-FGSM, TI-FGSM | No internal access; query or transfer | Cross-model gradient transferability | Commercial APIs, autonomous vision |
| Score-Based Query Attacks | ZOO, NES-PGD, Square Attack | Output probability scores only | Finite-difference gradient estimation | Cloud ML APIs, medical imaging |
| Decision-Based Attacks | Boundary Attack, HopSkipJump | Hard-label predictions only | Geometric boundary exploration | Robust black-box industrial systems |
| Universal Perturbations | UAP, GAP, CD-UAP | Image-agnostic single perturbation | Dominant singular vectors of Jacobian | Batch surveillance, deployment-scale |
| Physical-World Attacks | Adversarial patches, eyeglasses | Real sensor pipeline | Spatial transformations & lighting | Autonomous driving, face recognition |
| Backdoor / Trojan Attacks | BadNets, TrojanNN, ISSBA | Training-time data poisoning | Shortcut feature learning in DNNs | Federated learning, supply-chain ML |
| Prompt Injection / NLP Attacks | TextFooler, BERT-Attack, GCG | Token-level or embedding-level access | Discrete input sensitivity of language models | LLMs, sentiment analysis, NLU |
| Semantic & Geometric Attacks | ColorFool, Spatial Transform Attacks | Perceptual model of human vision | Semantic feature manifolds | Video surveillance, generative models |
| Model Extraction & Inversion | Knockoff Nets, MIA, Attribute Infer. | Query-only or gradient leakage | Decision boundary memorisation | Privacy-sensitive healthcare & finance |

Table 2. Adversarial Defense Mechanisms: Techniques, Deployment Stage, Robustness Guarantees, and Open Challenges

| Defense Strategy | Key Techniques | Stage of Deployment | Robustness Guarantees | Limitations & Open Challenges |
|---|---|---|---|---|
| Adversarial Training | PGD-AT, TRADES, MART, AWP | Training phase | Empirical robustness on seen attacks | High compute cost; robust-accuracy trade-off |

| | | | | |
|---|---|---|---|---|
| Certified Defenses | Randomized Smoothing, IBP, CROWN | Training + inference | Provable l-p ball guarantees | Scales poorly to large networks & datasets |
| Input Preprocessing | Feature squeezing, JPEG compression, bit-depth reduction | Pre-inference pipeline | Attack-surface reduction | Adaptive attacks bypass most preprocessors |
| Adversarial Purification | Diffusion-based purification, PixelDefend | Inference-time denoising | Stochastic denoising disrupts perturbations | Slow inference; may alter clean semantics |
| Detection-Based Defenses | LID, Mahalanobis detector, Deep KNN | Inference filtering layer | Identifies OOD adversarial inputs | High false positive rate; evadable |
| Ensemble & Diversity Methods | ADP, DVERGE, random subspace ensembles | Model architecture | Reduces transfer across sub-models | Increased inference cost; limited coverage |
| Denoised Smoothing | SmoothAdv + denoiser, DiffPure+RS | Inference with smoothing wrapper | Certifiable + empirical hybrid guarantees | Tight noise–accuracy trade-off |
| Architectural Robustness | Vision Transformers, Lipschitz networks, normalizing flows | Design time | Inherent geometric stability | ViTs vulnerable to texture-based attacks |
| Robust Feature Learning | Feature denoising, invariant risk minimization | Training feature extractor | Domain-invariant representation learning | Requires diverse training distributions |
| Game-Theoretic Defenses | Min-max robust optimization, Stackelberg games | Training with adversarial agent | Nash equilibrium robustness bounds | Computationally intensive; non-stationary adv. |

## 6. Cross-Cutting Themes/Implications.

### 6.1 The Trade-Off between Robustness and Accuracy.

The conceptual nature of seemingly fundamental trade-off between clean accuracy and adversarial robustness is likely the most repeatedly observed empirical result of adversarial machine learning, and one of the most theoretically discussed. In nearly all benchmarks, datasets, and architectures under investigation, enhancing adversarial robustness by using either of the defences mentioned in Section 4, is followed by a

decrease in clean test accuracy. One of the earliest theoretical accounts of this trade-off was given by Tsipras et al. (2019), who modeled a synthetic data distribution where a Bayes-optimal adversarial robust classifier must have lower clean accuracy than an equivalent Bayes-optimal standard classifier, and vice versa, as well as proved empirically that adversarial trained models learn qualitatively different features than standard models - an interesting implication that can be applied to model interpretability. Formalisation The trade-off has been formalised by Zhang et al. (2019), who found that robust error is at most the sum of natural error and some error term at the boundary, which needs to be reduced (to improve robustness) at the cost of natural error in a large model of distribution. Much more recent research has challenged this trade-off as being intrinsic or an artifact of small training sets, and has shown the large-scale pre-training on unlabelled data then fine-tuning with adversarial learning can substantially help close the gap enough to have reason to be optimistic that with sufficient data the robustness-accuracy trade-off can be significantly overcome.

### 6.2 Transferability and Implication To Security auditing.

It has far-reaching practical consequences because adversarial example transferability is applicable to models and architectures beyond the scholarly examination of model robustness. Security auditing In security auditing a transferable model means that an attacker who has access to openly available data can use the model to analyze the adversarial authenticity of a black-box system in an indirect way- a situation that is directly applicable to testing the safety of commercial ML APIs and hypercomputer deployments. Theoretical insights into transferability have made significant progress over the past several years: Ilyas et al. (2019) suggested the most popular hypothesis, according to which adversarial examples are non-robust features, which are statistically predictive but humanly inaccessible directions in the input space and shared by models that are trained on the same data distribution, thus links adversarial perturbations tapping into one model into activating adversarial perturbations tapping into another one. This theory replaces adversarial vulnerability as a defect in particular models with adversarial vulnerability being a property of the statistical geometry of natural training distributions, indicating that to get rid of adversarial vulnerability, one has to learn representations that are robustly non-vulgar features nonetheless.

### 6.3 Adversarial Defenses in High Stakes Domains.

The consequences of an adversarial vulnerability also do not admit of a uniform analysis: a misclassification due to an adversarial example in a consumer application tagging images is inconvenient, but an attack of the same type targeting a medical imaging AI,

an autonomous vehicle perception stack or a fraud detector model can have life changing or even disastrous impacts.

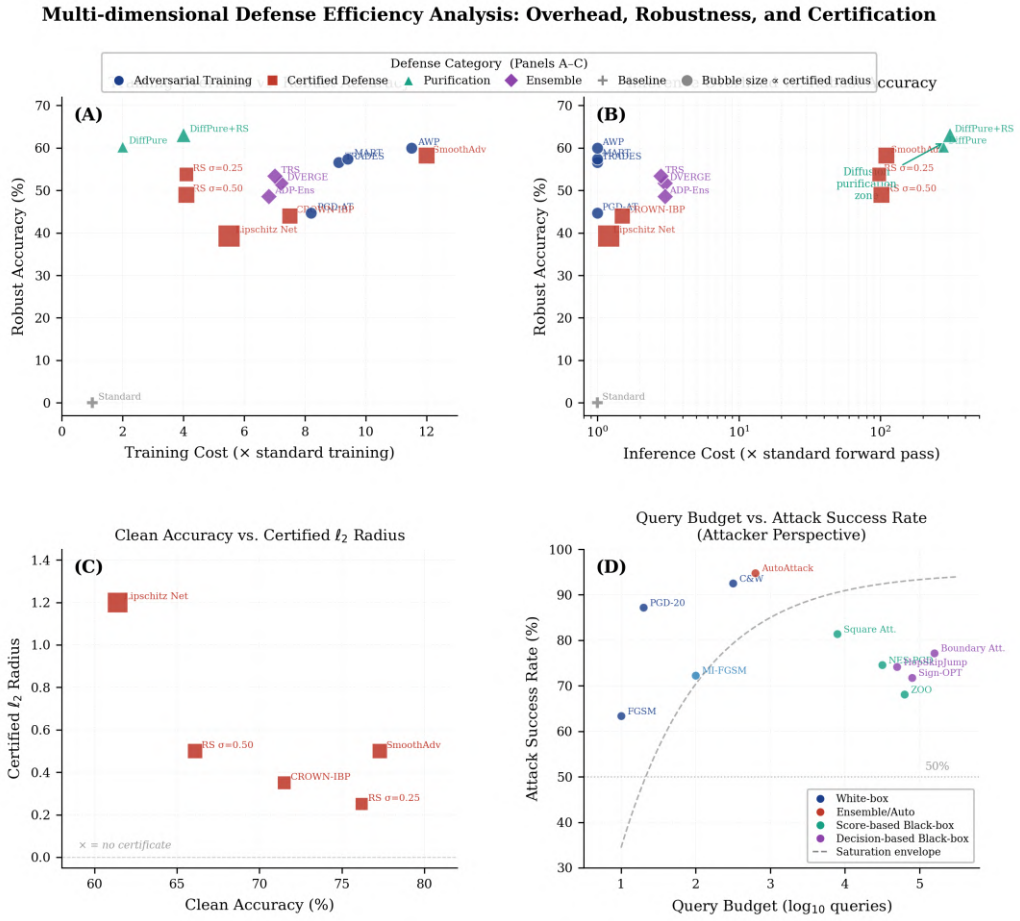


Fig.4 Defense Efficiency Map (2×2 Pairwise Grid)

Fig.4 shows A four-panel pairwise scatter grid examining adversarial defenses across multiple quantitative dimensions simultaneously: (A) training overhead vs. robust accuracy, (B) inference overhead (log scale) vs. robust accuracy — which isolates diffusion-based purification as a high-cost outlier — (C) clean accuracy vs. certified $\ell_2$ radius with iso-efficiency contours, and (D) an attacker-perspective view of query budget vs. attack success rate with a diminishing-returns saturation envelope. Bubble sizes encode certified radius where applicable.

As a subfield of medical imaging, peer-reviewed research demonstrating adversarial attacks on deep learning diagnostic systems have been established against models to solve skin lesion classification, diabetic retinopathy screening, histopathology image analysis, and chest X-ray interpretation models, and the success rate in attacks is more

than 90% in most studies under realistic threat models. The regulatory impact is also significant: the US Food and Drug Administration (FDA) and European Medicines Agency (EMA) are already securing that manufacturers of medical devices relying on AI start disclosing adversarial robustness tests as part of pre-market authorisation applications, and the risk-based system of classification requirements of the AI under the EU AI Act puts an increased pressure on the transparency and strength of AI systems used in high-risk industries. Physical-world adversarial patches In the field of autonomous vehicle, object detection and semantic segmentation systems are adversarial attacked, including in physical-world adversarial patches, LiDAR spoofing, and camera sensor attacks, they are an active topic of research and security engineering interest, and some automotive OEMs are now explicitly using adversarial robustness testing as a sub-requirement of the ISO 21434 cybersecurity engineering specifications.

### 6.4 Novel Frontiers: Multimodal, Generative and Embodied AI.

The boundary of adversarial machine learning study is swiftly broadening to the multimodal foundation models, generative AI frameworks, and the embodied AI agents. Vision-language models Adversarial attacks on models, including CLIP, LLaVA, GPT-4V, have shown that adversarial images can prompt multimodal models to produce specific false captions, misunderstand document content, or hallucinate specific objects-reveal new attack surfaces in AI-based document processing, scalable visual question answering, and embodied post-processing instruction following. Watermark evasion attacks, concept erasure attacks, and membership inference attacks have been established to be possible on text-to-image generative models, such as DALL-E, Stable Diffusion, and Midjourney. In the case of embodied AI systems, or robots and autonomous agents that act on the physical world based on deep learning perception and policy models, the cascading of adversarial adversity attacks on the perception module through the decision-making pipeline may be qualitatively different and worse than the isolated scenario of misclassification, which is the focus of a growing literature on adversarially robust reinforcement learning and robust model predictive control.

## 7. Open Research Directions and Conclusions

Machine learning Adversarial machine learning is an active area of study, and both theoretical and empirical challenges are considered open. One of the most urgent theoretical issues is whether the robustness-accuracy trade-off is a fundamental one or it is rather the constraint of the existing training paradigms and data magnitudes. It has been demonstrated recently in the context of self-supervised pre-training that the proposed trade-off is not necessarily irreducible; nevertheless, a full theoretical

characterisation of the circumstances under which the trade-off can be addressed, i.e. randomness of data distribution, capacity of the trained model, and training goals, is a high stakes unsolved problem. The problem of which attack threat model should be optimised against is also unresolved: norm-bounded perturbations are convenient to deal with technically, but are not an accurate model of real-world adversarial capabilities, which embrace arbitrary semantic transformations as well as composite, multi-attack, and time-correlated adversarial sequences of video or time-series. Insular growth of adversarial robustness assessments, either threat-model-independent (or realistic adversarial) aiming as a research direction, is significant.

Within the empirical front, the AutoAttack benchmark proposed by Croce and Hein (2020) has made significant progress by making adversarial robustness assessment far more reliable by not based on individual attack techniques, which might not be optimal in following particular defences, but an ensemble of complementary attacks, such as the parameter-free Auto-PGD, the FAB attack, and the Square Attack, to an empirical regimen of evaluation that is parameter-free and inadvisable to game. The RobustBench leaderboard that quantifies adversarial robustness testing across methods, architectures and datasets has played a significant role in offering reproducible benchmarks as well as detailing that most claims to improvements in adversarial defence fail to pass testing against the complete AutoAttack collection. Such contributions to infrastructure are scientifically significant as improvements in methodology: unless there are good assessment protocols, the improvement of adversarial robustness cannot be assessed with certainty. The future efforts should also focus on creating evaluation infrastructure of new frontiers of adversarial machine learning in multimodal, generative, and embodied systems where even the standard norm-bounded evaluation protocols cannot be applied and new evaluation frameworks need to be constructed out of first principles.

Summing up, adversarial machine learning has become more than just a surprising empirical finding in the vulnerability of image classifiers, to an established subfield that tackled basic questions on the geometry of learned representations, the effects of the certifiability of robustness, the security of large-scale deployed AI systems, and the reliability of AI in high-stakes applications. The direction suggests within the field has been one of a successive dialectic of ever more advanced attacks, and ever more principled defences, which develop each other--a trend which may be likened to the arms race of evolution which all security areas resemble. It has already made significant progress: certified robustness methods are now making provable guarantees about modestly-scaled networks; adversarial training variants are getting robust accuracy on CIFAR-10 that would have sounded impossibly high a decade ago; and adversarial robustness is now being considered a first-class design goal by the broader deep learning community instead of an optional safety measure. The difference between the levels of robustness that theoretically ought to have been attained and those that are achieved with

production-scale architectures and datasets, however, is still large and experiments of adversarial robustness are only starting to be extended to all of the variety of AI modalities and deployment conditions today. To address this gap it will in the long term be necessary to bring together interdisciplinary efforts that include theory of machine learning, optimisation, theory of cryptography, theory of formal verification, human-computer interaction, and theoretical theory of domain-specific safety engineering-a challenge that is as important as its goal.

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 8: Evaluating Trustworthy Artificial Intelligence: Metrics, Benchmarks, and Experimental Protocols

## Abstract

One of the most urgent and thought-provoking issues in the modern field of artificial intelligence is the assessment of trustworthiness of such systems as deep learning. With the introduction of neural networks into high-stakes paths, both clinical diagnostics and autonomous transportation, as well as financial risk analysis and criminal justice, the characterization of how to holistically and effectively examine their reliability, fairness, robustness, and uncertainty has left the theoretical fringes of both scholarly interest and regulatory debate and is at the forefront of them. This chapter provides a more organized and critical analysis of the measures, standards, and testing frameworks to have become the key tools of reliable AI testing. We see the development of the evaluation paradigm of simple holdout accuracy to multidimensional, adversarially educated, and distribution wise evaluation structures. We consider the conceptual basis as well as the technical constraints of calibration qualities, adversarial evaluation metrics, fidelity evaluations, out of distribution generalization assessments, and explainability testing methodologies. We also cover new attempts at standardization, the issue of benchmark saturation and overfitting, the increasing need to have evaluation frameworks that reflect operational conditions in the real world, and not in a laboratory. It ends with a prospective synthesis of unresolved challenges and an invitation to principled and community orientated benchmark governance in the service of the truly trustworthy artificial intelligence.

## 1 Introduction: The Call of Vigorous Trustworthy AI Review.

Implementing the deep-learning systems into the actual sociotechnical situations has generated a paradigm shift as to what constitutes a good model performance. Until relatively recently, machine learning benchmarking evaluation was a relatively simple operation: participants would divide an labeled dataset in training and test subsets, and

train a model, and report classification accuracy or mean squared error on the held-out set. This systems theoretically analytically clear but computationally tractable paradigm is based on a set of assumptions violated systematically in practice. The distributions in the real world vary in a continuous manner. Adversarial actors attempt to influence the inputs with the intention of enforcing inaccurate outputs. The features of the training data that represent sensitive information transfer discriminatory patterns to model predictions. And model confidence scores, often naively construed to be probabilities, are often not equal to observed empirical frequencies, in a way that led to human distrust and undermined the quality of decisions.

The rise of reliable AI as a subdiscipline of its own is a manifestation of the shared understanding that a model with a ninety-five percent accuracy on a benchmark can still have disastrous results, be unfair or erratic when exposed to the real world. As the concept is defined in the modern literature, the notion of trustworthiness is not a unitary property, and instead a composite concept that involves at least the following dimensions: technical robustness and stability to natural and adversarial perturbations; and calibrated uncertainty measures that allow downstream consumers to weigh the output of models appropriately; and algorithmic equity across demographic subpopulations; explainability and understandability of the reasoning mechanism; protection of privacy in connection with training data; and rich generalization outside the support of the training distribution. They all require special evaluation methodology and the interactions between them bring in more complexity that simple naive single metric evaluation paradigms are poorly placed to capture.

The consequences of poor assessment are not scholarly only. The inaccurate use of the wrongly prepared risk-scoring model in a healthcare facility can postpone or withhold life-saving measures. A self-driving perception system adversarially vulnerable can turn out to be disastrous. The results of a facial recognition system tested on majority-population benchmark data sets might be disastrous in the face of the underrepresented population groups. They are not just hypothetical situations, these are the reported cases which have come to bear the scrutiny of the regulating heights, the wrath of the people and in a few cases incurred legal consequences. The AI Incident Database, a list of actual artificial intelligence failures in practice, lists a number of hundreds of them; each of which is an evaluation failure at some point in the development lifecycle. It is against this background, that the design and the adoption of rigorous comprehensive, and deployment realistic evaluation frameworks, forms a scientific priority and an ethical requirement.

The chapter is structured in the following way. In section 7.2, the author explores the theoretical foundation and practical application of the metrics of trustworthiness on the key dimensions of evaluation. Section 7.3 is a survey on the landscape of benchmark datasets and standardized evaluation suites, and especially the design reasoning,

contributions of the empirical perspective, and limitations of the empirical perspective are being nationalized. Section 7.4 talks about protocols that make experimental use, such as train-test split and adversarial evaluation pipelines, and new practices in evaluating distributional robustness. The meta-problem of benchmark overfitting and, therefore, dynamic adaptive-evaluation frameworks is considered in section 7.5. There is a synopsis of open issues and future perspectives given in Section 7.6; and finally a couple of extensive summary tables given at the end of the chapter.

## 2 Trustworthy AI Metrics: Theoretical and Operationalization Empirical.

The paradigm known as Beyond Accuracy is a multidimensional evaluation paradigm.

The insufficiency of scalar accuracy as a measure of evaluating trustworthy AI systems has been abundantly recorded throughout the literature, but the accuracy-focused leaderboards and journalism continue to have significant gravitational force on the discipline. That accuracy is uninformative is not the basic issue but, rather, that it summarizes information, in a manner that confounds especially those aspects of model behavior that are of the greatest concern to reliable application. A classifier that classifies a dataset divided by a ninety percent majority at ninety-eight percent accuracy has been challenged with the trivial solution; predict the majority class always; the accuracy of any such answer is beyond reproach, but its practical use is nonexistent. In a similar way, a highly average fitting model can also get that aggregate by radically disproportionate representation of performance by demographic subgroups, geographic regions or modalities of the input, the production of a misleading impression of overall competence.

There is an academic consensus that has been placed upon a multidimensional understanding of evaluation that includes dimensions whereby accuracy measures and metrics are required, but are not exclusive. Calibration measures determine the agreement of model confidence and empirical validity. Robustness metrics are measures of the sensitivity of model predictions to perturbed inputs of different kinds, such as Gaussian noise, adversarial examples, distribution shift or domain shifts. Fairness measures are normative measures of fair treatment between populations. Matrices of quantification of uncertainty examine the quality of the predictive distributions such as the sharpness, coverage and reliability of the predictive distributions. The metrics of explainability are aimed at quantifying the extent to which post-hoc attribution procedures are indicative of the internal decision process. And privacy measures are a measure of the extent to which the model outputs can be used to inform training data. Collectively these dimensions make up the evaluation terrain of credible AI, and which any serious empirical contribution to the field has to wrestle specifically with its subset.

### 2.1 Metrics of Calibration and Uncertainty Quantification.

Calibration takes one of the most prominent places among the metrics, which have been under best consideration in the post-accuracy evaluation paradigm. An economically fit model is described to be well-calibrated when the probability of a successful prediction as to a particular success is measured as its prediction and is equal to the actual frequency of that success. In the formal sense, perfect calibration means that given any confidence value, p, the proportion of predictions made with that confidence level that are successfully made is p. The Expected Calibration Error (ECE) operationalizes this criterion by segmenting the range of confidence into M equal-width bins, and then averaging the absolute biggest gap between the average confidence and the average accuracy in each of the bins, and take an average across bins with a frequency-weighted mean. Although ECE is now the de facto standard metric of calibration, it has been found to be limited in a number of ways: it does not punish the lack of sharpness, and it can be gamed by models that do not distribute predicted values over the range of the confidence scale.

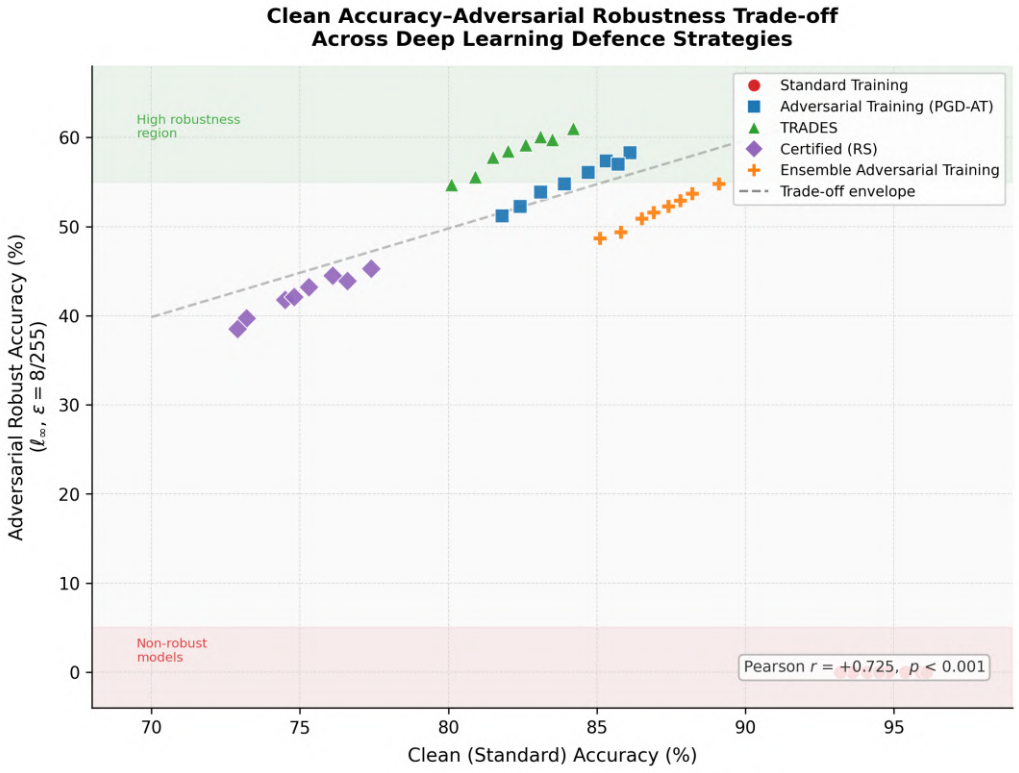


Fig.1 Clean Accuracy vs. Adversarial Robustness Trade-off (Pairwise Scatter)

Fig.1 shows A pairwise scatter plot comparing clean accuracy against adversarial robust accuracy (ℓ∞, ε=8/255) across five defence strategies — Standard Training, PGD-AT,

TRADES, Certified Randomised Smoothing, and Ensemble Adversarial Training. Includes a fitted exponential trade-off envelope, Pearson r annotation ($r = +0.725$, $p < 0.001$), and shaded risk regions. Directly visualises the core accuracy–robustness trade-off discussed in §7.2.3.

More advanced metrics of calibration have been suggested to overcome this limitation. The Adaptive ECE solves the bin-boundary sensitivity issue by employing quantile based binning over uniform binning making sure that all the bins have equal count of predictions. The Continuous Calibration Error conducts a bio swap of discrete binning by means of kernel density estimation generating a flatter and bin-free calibration rating. Reliability diagrams, a visual representation of average confidence versus average accuracy by bin, have become a complement to scalar calibration scores and are also becoming increasingly used in conjunction with ECE in empirical literature. In addition to calibration, proper scoring rules, like the Negative Log-Likelihood (NLL) and the Brier Score, are used to measure the quality of full predictive distributions in a mathematically principled manner they both penalize accuracy and calibration. The difference between the notion of calibration (precision of confidence levels) and sharpness (dispersion of the predictive distribution) has been rationalised by Gneiting and Raftery (2007) in a scheme that still is often used in setting up benchmarks in the design of uncertainty quantification.

The use of Bayesian deep learning models, such as Monte Carlo Dropout or Deep Ensembles and more recent models, such as SNGP (Spectral-Normalized Gaussian Process), and Evidential Deep Learning, has led to other measures of the quality of epistemic and aleatoric uncertainty decomposition. Of particular importance to out-of-distribution detection is epistemic uncertainty, a phenomenon caused by the uncertainty in the model parameters and that tends to go down with the addition of training data to the model whereby in-distribution inputs to the model should have low uncertainty and out-of-distribution should have high uncertainty. The calculation of the weighted average between in-distribution and out-of-distribution uncertainty scores (AUROC) has become a common measure of such who has obtained scores, but the interpretation of results requires heavily based on the selection of OOD dataset and the definition of what should be considered in-distribution data.

### 2.3 Metrics of Adversarial Robustness

Adversarial robustness testing has significantly developed since the pioneering Szegedy et al. (2013) and Goodfellow et al. (2014) works that established that it was consistent to overt small input perturbations to make state-of-the-art classifiers issue false outputs with high confidence. The Fast Gradient Sign Method (FGSM) and the Basic Iterative Method (BIM) are the two attacks, which are computationally inexpensive and offer

weak vectors of model vulnerability, which dominated early robustness tests. The further evolution of the Carlini-Wagner (C&W) type of attack, the Projected Gradient Descent (PGD) type of attack, created by Madry et al. (2018), and finally, the AutoAttack framework developed by Croce and Hein (2020) set more strict requirements on the evaluation of adversarial robustness. The most used benchmark-grade adversarial evaluation methodology is AutoAttack, a Multi-attack ensemble combining four complementary attacks, including a parameter-free version of PGD and two black-box attacks, and it is precisely due to its multi-attack ensemble being significantly harder to defeat by non-black-box adversarial evaluation methods, which mostly overfit on a specific attack algorithm.

This differs empirical robustness (quantified in terms of accuracy against certain attacks), and certified robustness (in which a formal proof shows that the predictions will be stable in a norm-bounded area about each input), is perhaps one of the most significant methodological splits of the subject. Proven robustness methods, such as randomized smoothing and mixed-integer linear programming verification techniques, are certifiable, though commonly at the expense of significantly decreased clean accuracy and fewer computational resources. The certified accuracy at radius r is the probability of the test inputs such that the model prediction is provable to be stable under all perturbations in an L2 or Linf ball with radius r and this is becoming an increasingly common measure in rigorous assessments alongside empirical adversarial accuracy. The accuracy-robustness trade-off, which is widely reported in the literature, is one of the major open problems in the area: nearly all the known algorithms to enhance adversarial robustness are associated with negative trade-offs in clean accuracy, and even the strength of the trade-off is a significant metric by itself.

### 2.4 Fairness Measures and Tensions Inherent in Fairness

With regards to the AI literature, algorithmic fairness has produced one of the most prolific and anecdotal evaluations methodology. The spread of the formal criteria of fairness, such as demographic parity, equalized odds, individual fairness, counter faction fairness, and calibration within groups, among many others, are indicative of both the multidimensional nature of the notion of fairness and the inability to meet all the criteria in most realistic scenarios. Formal proofs of the impossibility theorems of Chouldechova (2017) and Kleinberg et al. (2017) that in the non-degenerate situation when groups have different base rates, both group calibration and equalized odds are impossible, cannot both be met concurrently. This outcome has far-reaching consequences on the measure of fairness: there is no single metric that one should be fair, and one of the metrics of fairness to maximize and assess is the structure of a normative choice that needs to be made explicit.

Broadly, fairness measurement at the deep learning systems has generally been applied in stages: as the first stage, it identifies protected attributes of interest such as race, gender, age, socioeconomic status, or proxies to it, followed by calculating disaggregated performance measures within subgroups based on these attributes. The difference between the groups of different true Positive rates and the difference between the groups of different false Positive rates calculated as a maximum of the difference gives a fairness criterion at the group level that is sensitive to both errors. It is the Equalized Odds Difference, which is the maximum of the difference between the groups of the different true Positive rates plus the maximum of the difference between the groups of different false Positive rates. Demographic Parity Ratio this is a ratio calculation of the positive rates of prediction in a group, which is used when the result itself is a matter of concern, or it is required by legal doctrine disparate impact analysis. The Predictive Parity criterion that demands all groups to have equal precision conforms to the intuitive view that individuals in other groups with the same score on the model should equally have the same chance of the target outcome. Practitioners and regulators are becoming more and more demanding of evaluation reports to capture a comprehensive fairness audit which reports the performance across all applicable subgroup intersections, and not just on single demographic dimensions as the intersectional fairness proposed by Buolamwini and Gebru (2018).

### 2.5 Interpretability and Attribution Metrics

The interpretability of a model is the perhaps the epistemologically hardest problem in the responsible evaluation of AI, since it involves attempting to measure an attribute, faithfulness of explanation, which is formulated in terms of the hidden truth about the internal reasoning of a model. Such post-hoc explanation algorithms as SHAP, LIME, Grad-CAM and Integrated Gradients produce attribution scores that are purportedly used to infer which input features were the most important in a given prediction, and they can be heavily biased by the inductive biases in the explanation algorithm and can be not representative of how the model actually reaches the answer. These measures include the Area Over the Perturbation Curve (AOPC), which tries to evaluate faithfulness estimating how adding successively the most salient features by the explanation diminishes model performance with faithful explanations predicted to have a steeper curve. This metric in turn is confounded by the manner in which replacement value of masked features is adopted so as to include a distributional shift which in itself independently impacts on model performance.

More recent efforts have led to less pragmatic interpretability assessment via the prism of model ground truth. With synthetic benchmark datasets whose generative mechanisms are known i.e. the ground truth causal graph between features and labels

defined, it is possible to evaluate whether an explanation method identifies features that are causally relevant. Hooker et al. (2019) present a more rigorous form of perturbation-based faithfulness assessment as ROAR (Remove and Retrain): they retrain the model after removing the features to remove the confounding the effect of out-of-distribution examples. The problem of human-based assessment whereby domain experts can judge the plausibility and usefulness of model explanations to realistic situation decision-making processes has also been on the upward trend as a complement to automated measures, aware that the explanations might be computationally faithful but not useful to human decision-makers.

## 3 Trustworthy AI Benchmarks: Design, Contribution and Critique

### 3.1 The Architecture of a Trustful Artificial Intelligence Benchmark

A designed reliable AI benchmark has several simultaneous purposes: it offers a standardized evaluation protocol that allows the achievement of fair comparison of the methods and research groups; it reveals particular failure modes or evaluation dimensions that standard datasets fail to probe in a satisfactory way; and it propagates a common empirical vocabulary within which the research community can share understanding of progress. Such a benchmark design entails successive normative and technical decisions: what to evaluate, how to create or maintain data, what threat models and fairness standards to use, what evaluation metrics to use, and how to govern the process of reporting results, who is allowed to submit results. All these options embed assumptions as to what credible AI needs, and history of benchmarking shows poorly thought benchmark tests may misguide a research initiative at a huge cost.

Although the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is not a reliable AI benchmark currently and can serve as a valuable historical case study. It has a clean, balanced and curated structure, which allowed in dramatic improvement in the accuracy of classifying the images, but it also introduced an evaluation monoculture where the models have been trained on a particular image distribution and evaluation protocol that increasingly diverged with how they would work in practice. This was later identified as the extent that reported performance was due to overfitting to dataset-specific features instead of true generalization downstream via a series of studies by Recht et al. (2019), who built a new test set in the same ImageNet protocol reported set and reported significant decline in accuracy across all models tested on the new test set. This result, which is currently much discussed as the driving factor in creating frameworks of evaluation more resistant to gaming, exemplifies a broader rule: standards

that become fixed objectives become liable to both intentional and unintentional cases of gaming.

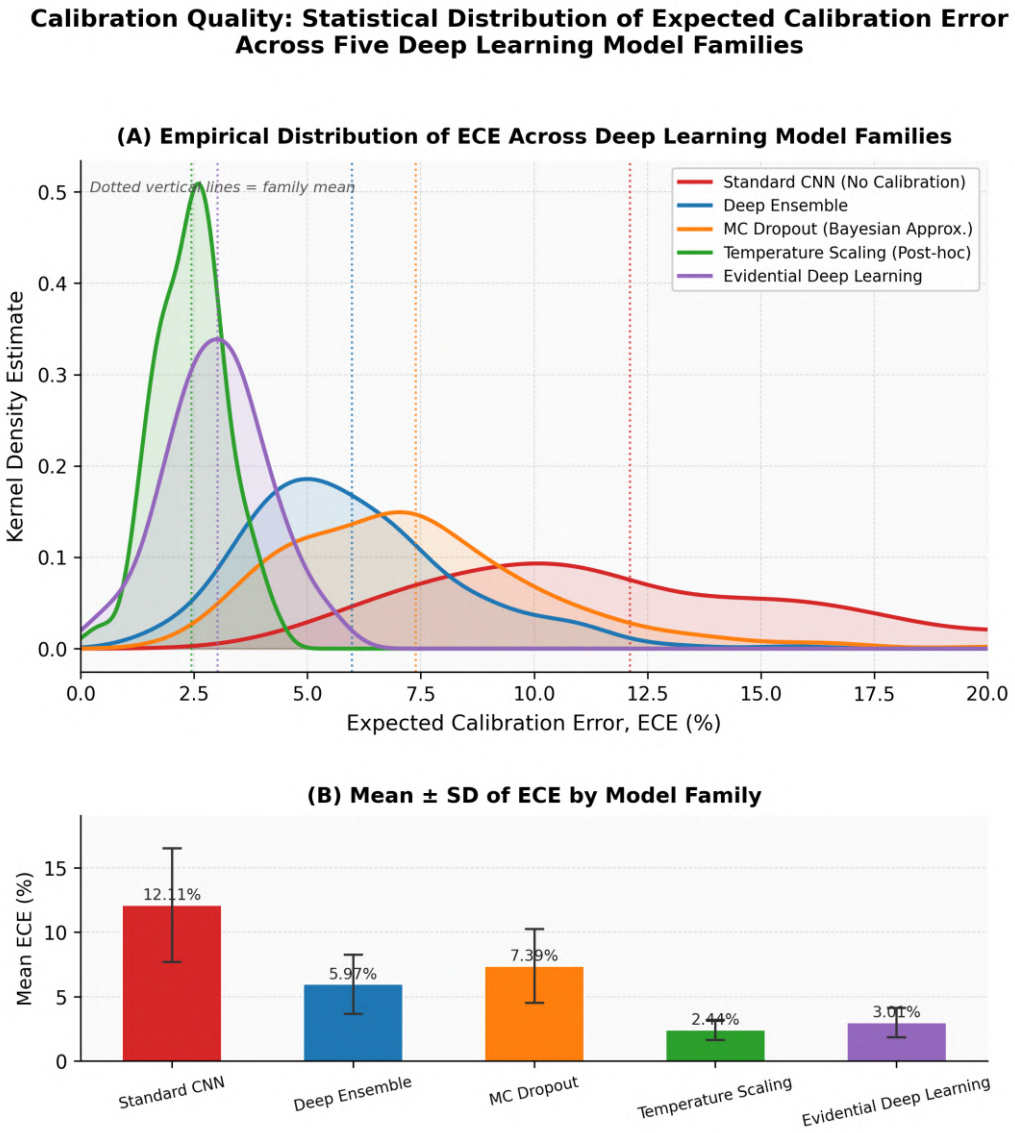


Fig.2 Statistical Distribution of ECE Across Model Families (KDE + Bar)

Fig. 2 is A dual-panel figure with overlapping kernel density curves (Panel A) and mean ± SD bar chart (Panel B) comparing Expected Calibration Error distributions across five model families: Standard CNNs, Deep Ensembles, MC Dropout, Temperature Scaling, and Evidential Deep Learning. Dotted vertical lines mark family means. Illustrates the calibration quality hierarchy discussed in §7.2.2.

### 3.2 Benchmarks on Robustness: RobustBench, WILDS and ImageNet-C.

One of the most effective contributions to standardized adversarial robustness evaluation is The RobustBench benchmark, presented by Croce et al. (2021). RobustBench allows a transparent leaderboard to be maintained publicly with predetermined AutoAttack model checkpoints to overcome the reproducibility crisis which had plagued robustness assertions in earlier studies, most of which were found to indicate artificially inflated images of robustness. It includes the Linfty and L2 threat models and tests on CIFAR-10, and CIFAR-100 and ImageNet, offering a mechanized perspective on the dependence on good scale with the complexity of the data set. One of the insights made possible by RobustBench is the consistent trade-off between accuracy and robustness: the most robust models on RobustBench are achieving significantly lower clean accuracy than any of the standard-training models, and the boundary of this trade-off has been improved only gradually despite years of research.

WILDS benchmark, created by Koh et al. (2021), considers the complementary aspect of robustness, i.e. the capacity of models to provide the generalization across the distribution shifts which occur naturally. Unlike adversarial benchmark methods, which evaluate the performance when the adversary introduces worst-case perturbations, WILDS constructs ten real-world experiments and these dataset shifts happen due to authentic ecologically relevant factors- variations in hospital, geographical area, demographic structure, or even time. This is a significant point of difference in that approaches enhancing adversarial, free of distributional robustness and the other way around are important to understand that they may be tackling entirely different aspects of how brittle models are. WILDS has shown that domain-invariant learning algorithms despite their theoretical justification tend not to achieve consistent performance relative to empirical risk minimization with respect to out-of-distribution test sets, which is of great relevance to the development of systems design that are pragmatically robust.

ImageNet-C and ImageNet-P, introduced by Hendrycks and Dietterich (2019), are corruption and perturbation robustness benchmarks based on a standardized set of fifteen popular corruptions, such as Gaussian noise, motion blur, compression artifacts, and weather effects, and a set of five levels of severity. These metrics obtained as the Mean Corruption Error and the relative Mean Corruption Error based on these benchmarks have become common reporting practices in computer vision papers purporting to contribute to robustness, and allow the consistent comparison across hundreds of methods. It has also been applied to other modalities and tasks (ImageNet-R (rendition based distribution shift), ImageNet-A (adversarially filtered natural images) and ImageNet-Sketch), which can be considered an in-depth diagnostic for understanding which specific kinds of distributional variation a given model is and is not resistant to.

### 3.3 The Question of Holistic Evaluation and the flaws of a language model benchmark.

Model assessment has become one of the most hotly debated fields in the development of benchmarks, with such models as GPT-4, Claude, Gemini, and LLaMA postponing incredible abilities and equally incredible damage on the human being. Consisting of multitask evaluation tasks such as GLUE and SuperGLUE did provide an evaluation framework to test various types of linguistic understanding that have matured over years, but the benchmarks have quickly become saturated--modes of models that obtain near-human or superhuman aggregate performance within a few years of creation, but no proportional insight into the true logic of how these models reason or are to fail. It was hoped that the BIG-Bench collaborative project, containing more than two hundred assignments that covered mathematics, coding, social reasoning, and various language phenomena, could be sustained to be difficult until the predictable future both by having tasks of the size that the currently available models can handle and by charting performance curves as models increase in power.

Perhaps the most promising effort so far to operationalize a multidimensional assessment of language model trustworthiness, the HELM (Holistic Evaluation of Language Models) framework, created at Stanford represents an extremely ambitious attempt to do so. Comparing thirty models under forty two scenarios and seven fundamental metrics including accuracy, calibration, robustness, fairness, efficiency, toxicity, and bias, HELM offers a comparative analysis in a structured form that reveals the trade-offs that are opaque to the single-metric analysis. The major HELM results are that the high accuracy models on traditional benchmarks tend to be poorly calibrated, the measures of toxicity and bias differ widely across models on ostensibly similar training procedures, and that no model can be strongly predictive across all seven of the trust-relevant dimensions at once. The DecodingTrust benchmark, which is explicitly designed to investigate the trustworthiness of GPT-4 on a set of eight dimensions, such as privacy, fairness, adversarial robustness and stereotype susceptibility, illustrates once again that even models instructed on a particular topic or safety-trained are still susceptible to adversarial prompting and display quantifiable demographic biases, that the alignment fine-tuning is not as close to complete trustworthiness as it is reputed to be.

### 3.4 Multimodal Fairness, Privacy and new benchmarks.

The creation of fairness-specific criteria has sped up greatly after the recording of demographic inconsistencies in commercial facial analysis systems, the most notorious one being the Gender Shades report by Buolamwani and Gebru (2018), which observed error rates of dark-skinned female three times that of light-skinned men across three commercial facial analysis APIs. The FairFace and DemogPairs datasets were

constructed so as to permit systematic assessment of the systemic bias in any face analysis system, by using datasets with increased demographic balance in comparison with a previous bench-mark such as LFW. Fitzpatrick17k dataset of skin tone-labeled dermatological images, as well, have allowed the assessment of inequalities in fairness in the context of other medical imaging, and found that medical models trained on typical clinical datasets achieved significantly worse results on darker skin colors, a clinical interpretation with direct clinical implications, since the demographic differences in the outcome of skin cancer.

Privacy assessment standards are more modern, as the increased awareness of deep learning models storing training data in a way that provides membership inference and attribute inference attacks. Vision-language PrivacyLens benchmark, which judges the extent to which multimodal models disclose sensitive personal data to vision images encountered through inference despite the model has been fine-tuned via safety alignment, assesses how multimodal models reveal such sensitive data, such as approximate age, health conditions, and location. The machine unlearning evaluation protocol, which measures whether models whose training set has been nominally pruned nonetheless display the remnant of the influence of that data, has also become popular with privacy regulations like the right to erasure in the GDPR setting generating practical requirements of demonstrably forgettable capabilities in machine learning models. Together, these benchmarks indicate an expansion of the reputable AI assessment program to include all of the properties that constitute the existence of whether a deployed AI system can be trusted by the individuals and communities that it interacts with.

## 4 Experimental Procedures of Steel-Plate Reliable AI

### 4.1 Train-Test Protocol Standardization

Reliability of credible AI assessments is highly reliant on standardization of test elements at each phase of the assessment channel. Even apparently insignificant choices, such as using a random seed to divide data, the normalization of inputs, the temperature multiplier of using model logits to estimate calibration metrics, the specific adversarial attack implementation, or what should be considered a positive prediction for binary fairness analysis may effectively lead to changes in reported metrics that would make an achievement of new techniques seem pale in comparison. The relevance of standardized evaluation procedures has been established severally in benchmark experiments that remodel already published procedures under controlled environments and observe

significant variances with the results reported to them, not due to fabrication of data, but to the summing up to small implementation decisions.

Good AI experimental procedures guidance consists of preregistration of the evaluation procedures prior to observing results, publication of full implementation of the evaluation codebases and model weights, utilization of fixed random seeds and deterministic evaluation modes to ensure reproducibility and publication of the various runs with reasonable quantification of their uncertainty on the reported metrics. Several evaluation frameworks including evaluation of Hugging Face, the LM Evaluation Harness of EleutherAI, and the TorchMetrics have partly solved the standardization issue by offering reference implementations of common metrics, though the evaluation ecosystem is still highly fragmented, and thus prevents non-experimental, reproducible, and comparable research. New practice protocols such as model cards and labels on the nutrition of datasets, which are more communicative than technical measures, also provide standardization in the protocols as they dictate systematic documentation of evaluation conditions.

### 4.2 Pipelines of Adversarial Evaluation

Adversarial evaluation pipelines design includes a series of decisions that, jointly, are used to decide on the quality and informativeness of robustness testings. The choice of the threat model, such as the attack norm (L0, L1, L2, or Linf), the perturbation budget (epsilon) and whether the adversary access using the black-box or white-box is a limiting element, is the base decision on which all the other decisions should be consistent. White-box attacks, based on complete information on both the model parameters and gradients, give the strongest robustness test, but do not necessarily represent the access capability of a real-word adversary. Black-box attacks that solely act on model inputs and outputs are more realistic to deployment but offer weaker robustness estimates that can be deceptive on gradient obfuscated models. The white-box and black-box attacks of the AutoAttack framework is a principled solution to this tension, and offers an evaluation protocol which is practicably infeasible to win without adversarial capabilities unattainable in the real world.

In the case of natural language processing, a pipeline doing adversarial evaluation has to deal with discrete, non-gradient-based attacks which do not allow gradient-based attacks in the strict sense, and thus requires other approaches, such as token substitution, sentence paraphrasing, character-level perturbations, and semantically preserving transformations. Text adversarial attack pipelines offered by the TextFooler, BERT-Attack, and CLARE models and incorporated into NLP robustness testing systems are more sophisticated. The parameterisation of semantic similarity constraints, which in general require embedding distance or human judgment to ensure that adversarial

examples do not change the original sense, is also an added complication not available in the continuous-input case, and irregular applications of the parameterisation has been a key cause of incomparability among reported robustness performances.

### 4.3 Distribution Shift and Out-of-Distribution Evaluation Protocols

When conducting model generalization evaluation when there is a shift in distribution, one needs to be careful about the type and the extent of the shift as well as the correlation between the training and test distributions and the performance degradation metrics that is used to summarize the effect. The ImageNet-C standard language of corruption robustness testing introduces types of corruption unperceived in training at varying degree levels, and reports the performance as a function of severity, where it is possible to not only examine whether corruption affects models, but also to evaluate the manner in which performance is gracefully degraded as the corruption severity rises. This degradation curve offers more diagnostic data than a single scalar one and has become a reporting standard in publications to date on robustness.

Evaluation protocols of out-of-distribution detectors should deal with the inherent problem, which is that the notion of what is out-of-distribution is relative to a given in-distribution, and various selections of OOD dataset can yield radically different results in the relative ranking of detection procedures. Yang et al. (2022) conducted a systematic comparative study of OOD detectors across various OOD datasets and found out that no specific OOD detection method has prevailed in all conditions of evaluation, and reported values of the AUROC were highly sensitive to the choice of OOD dataset. This observation is the basis of OOD evaluation protocols which report performance via a diverse set of OOD datasets other than when a clear point of favorable measurement is chosen, and in this regard has served to inspire a set of established OOD evaluation suites like OpenOOD.

### 4.4 Benchmark Saturation, Overfitting and the Dynamism of Evaluation.

The issue of benchmark saturation and overfitting is one of the greatest meta level challenges facing the trustworthy AI evaluation community whereby the advancement of research on a benchmark is an optimization to benchmark specific properties rather than an actual advancement on the underlying capability of interest. This phenomenon is most clearly exemplified by the history of NLP benchmarking: GLUE was introduced in 2018 as a hard standard of multitask performance, and within two years, models had blended well-performed basis in aggregate GLUE scores, but had severe failures on the reasoning-based components of the benchmark where proper language comprehension is required. It was succeeded by SuperGLUE that was an even more tougher challenge

and even that started saturating around a similar period of time. The implication of this pattern is what will be referred to as a core dynamic, that is, the decrease in the difference between a benchmark achievability and actual achievability does not occur as the models become more genuinely capable, but due to the model learning to utilise benchmark-specific artifacts, biases in annotation patterns, or distributional properties of the evaluation data.

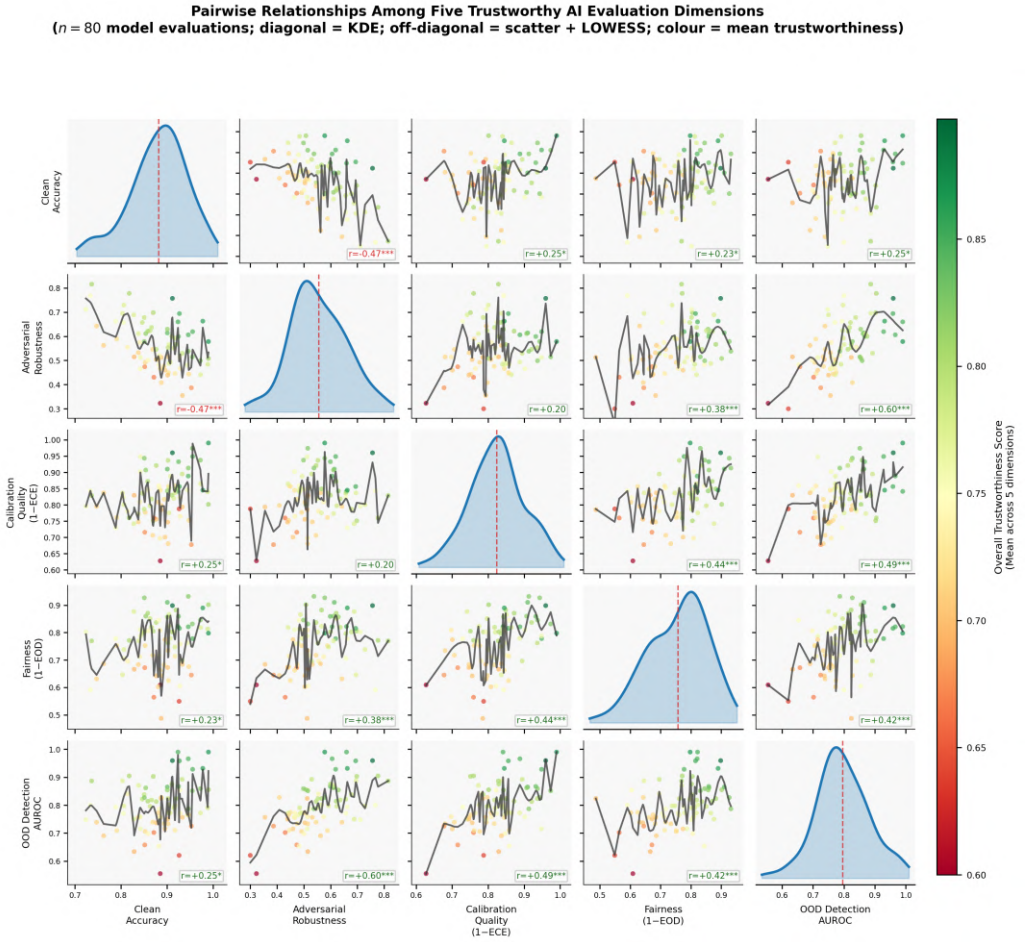


Fig.3 Pairwise Scatter Matrix (SPLOM) of Five Trustworthiness Dimensions

Above fig.3 shows A 5×5 scatter-plot matrix with marginal KDE distributions on the diagonal and pairwise scatter + LOWESS smoother curves on off-diagonal panels, for Clean Accuracy, Adversarial Robustness, Calibration Quality (1−ECE), Fairness (1−EOD), and OOD Detection AUROC. Points are coloured by overall mean trustworthiness score (RdYlGn scale). Every panel reports Pearson r with significance stars (*, **, ***). Supports the multidimensional evaluation argument of §7.2.1.

Table 7.1: Core Trustworthy AI Evaluation Metrics — Definitions, Domains, and Limitations

| Metric Category | Specific Metric | Definition / Computation | Applicable Domains | Limitations & Considerations |
|---|---|---|---|---|
| Accuracy & Performance | Top-k Accuracy | Fraction of samples where true label appears in top-k predictions | Image classification, NLP, retrieval | Misleading in class-imbalanced settings; ignores calibration |
| Accuracy & Performance | F1-Score (Macro/Micro) | Harmonic mean of precision and recall, averaged across classes | Text classification, medical diagnosis | Requires careful averaging strategy; sensitive to class distribution |
| Robustness | Certified Accuracy | Fraction of inputs for which predictions are provably stable under perturbation | Safety-critical AI, autonomous systems | Computationally expensive; scales poorly to large models |
| Robustness | Empirical Adversarial Accuracy | Model accuracy under iterative adversarial attacks (e.g., PGD, AutoAttack) | Computer vision, malware detection | Attack-dependent; may not reflect true worst-case robustness |
| Uncertainty Quantification | Expected Calibration Error (ECE) | Average difference between predicted confidence and empirical accuracy per bin | Probabilistic models, Bayesian NNs | Bin-width sensitive; does not capture sharpness or resolution |
| Uncertainty Quantification | Negative Log-Likelihood (NLL) | Measures probabilistic quality of predictive distributions over test data | Regression, classification, generative models | Penalizes overconfident wrong predictions; requires probabilistic output |
| Fairness | Equalized Odds Difference | Gap in true positive and false positive rates across demographic groups | Recidivism prediction, credit scoring | Trade-off with accuracy; definition varies by legal jurisdiction |
| Fairness | Demographic Parity Ratio | Ratio of positive prediction rates between protected and reference groups | Hiring algorithms, healthcare triage | May conflict with other fairness criteria (e.g., calibration) |
| Interpretability | AOPC (Area Over Perturbation Curve) | Measures explanation faithfulness via iterative feature masking and accuracy drop | Saliency methods, attribution techniques | Metric itself depends on a model forward pass; computationally intensive |
| Out-of-Distribution Detection | AUROC (OOD Detection) | Area under ROC curve separating in-distribution from OOD samples using model confidence | Deployment safety, open-world recognition | Assumes confidence as a proxy for in-distribution-ness; can fail silently |

| | | | | |
|---|---|---|---|---|
| Privacy | Membership Inference AUC | Discriminability of membership status of training samples from model outputs | Federated learning, healthcare AI | Attack-dependent; requires a shadow-model setup to estimate |
| Efficiency-Trustworthiness Trade-off | Robustness per FLOP | Certified or empirical adversarial accuracy normalized by inference compute cost | Edge AI, resource-constrained systems | Normalization baseline varies across hardware; non-standardized |

The idea of benchmark overfitting, which is analogous to the classical overfitting to a training set problem, is the effect where a model or research procedure is successively employed to fit onto a benchmark in a reiterative improvement procedure which utilizes information on the properties of the benchmark. Caused by random data contamination or not, the recursive operation of architectural, hyperparameter, and training process selection guided by benchmarking performance is also a type of meta-level optimization that overstated reported performance as compared to actual generalization. This issue is compounded with large language models, which are trained on large bodies (internet scale) of text that might contain the text of popular benchmark papers and their sample cases as well as even the answer keys to those cases, causing an acute data contamination issue, beyond the scope of existing train-test separation protocols.

The solutions suggested to benchmark saturation and overfitting are suggested on technical and governance levels. Technically, because dynamic benchmarks create and recreate evaluation instances using programmatic or adversarial algorithms, e.g. Dynabench, which creates evaluation instances by annotating the model-in-the-loop, will provide a moving target that is less likely to be saturated. The presence of held-out test sets which are never made public combined with secure evaluation servers which only report aggregate measures but do not reveal individual predictions restricts how much researchers can reverse-engineer and taste towards particular evaluation instances. Adaptive evaluation protocols with overlaying specific dimensions of diagnostic reports of capability, less than aggregate reports, present information more rich and interpretable on model strengths and weaknesses. On the governance side, benchmark transparency policies making training data composition and hyperparameter search procedures as well as any benchmark-specific tuning be disclosed are actively advanced as substantial research integrity.

### 4.5 Emerging Frontiers: Foundation Models, Multimodal Systems, and Real-World Deployment Evaluation

Large foundation models (trained on large multifaceted datasets and with the ability to quickly adapt to diverse downstream tasks by prompting or fine-tuning) have presented a new breed of evaluation issues that are not effectively discussed by existing reliable AI frameworks. The size of such models, as well as the generality of their capabilities implies that such a thorough evaluation will involve covering a large task space compared to the coverage that could be achieved by a single such benchmark. The HELM and BIG-Bench systems are initial phases in the direction of holistic assessment, but they still just do not reflect a whole set of abilities and malfunctions of models such as GPT-4 or Gemini Ultra. The interpretability issue is also qualitatively worse in the case of foundation models as in vivo mechanistic interpretability studies, attempting to understand how certain capabilities are produced by the computation implemented by specific computational circuits are in their infancy, and have severe scaling difficulties.

Multimodal AI systems, involving visual, linguistic and audio and possibly other sensorial modalities impose further testing complexities due to the interaction between modalities and the difficulty in specifying semantically identical perturbation across modality lines. Simple vision-language model, which properly captures an image, can be disastrously inaccurate on adversarial perturbations of the image that cannot be perceived as such by humans, and robustness in such a multimodal context can only be evaluated by attack and assessment protocols that consider cross-modal consistency. PrivacyLens benchmark and recent advances on adversarial multimodal evaluation have already started to overcome these problems, yet a holistic multimodal trustworthiness assessment framework is an ongoing research problem.

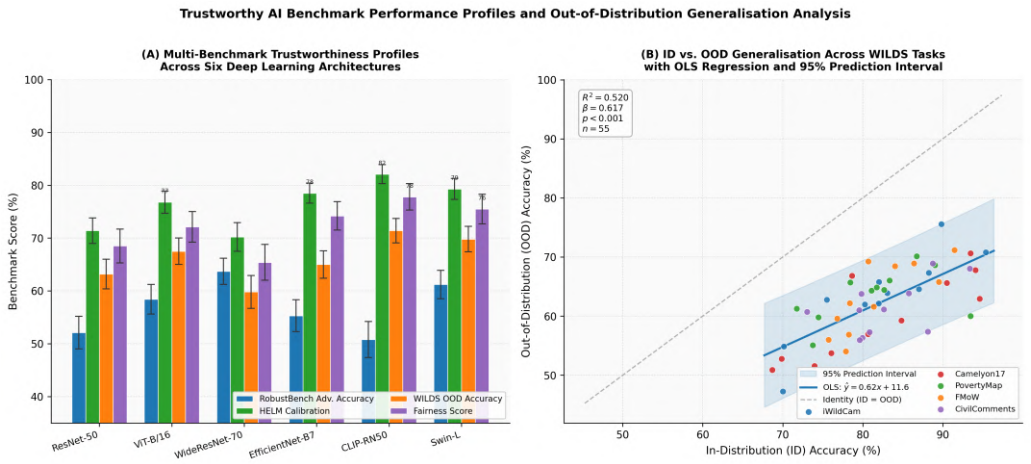


Fig.4 Multi-Benchmark Profiles & ID vs. OOD Generalisation Regression (Grouped Bar + Regression)

Fig.4 is A two-panel figure: Panel A is a grouped bar chart with 95% CI error bars comparing six architectures (ResNet-50, ViT-B/16, WideResNet-70, EfficientNet-B7, CLIP-RN50, Swin-L) across four benchmark scores (RobustBench, HELM, WILDS, Fairness). Panel B shows an OLS regression scatter plot of in-distribution vs. out-of-distribution accuracy across WILDS tasks, with a 95% prediction interval band, identity line, and full regression statistics ($R^2 = 0.52$, $\beta = 0.617$, $p < 0.001$). Directly supports §7.3.2's discussion of WILDS and multi-benchmark evaluation.

The real-world testing of AI systems under deployment conditions rather than one in a controlled laboratory setting has become a highly important frontier to credible AI research. The listed catalogues of deployed AI failures, such as the AI Incident Database, attest to the fact that benchmark performance is a very poor predictor of the reliability of a deployed AI system not through a lack of information but due to abstracting away the sociotechnical context: the organizational incentives, human-AI interaction pattern, edge case distribution, and feedback loop which make a deployed AI system behave in practice. Recent methods of deployment-realistic evaluation are red-teaming protocols where human adversaries systematically attempt to cause harmful or untrustworthy outputs out of deployed systems; shadow deployment evaluation, where candidate models are fed actual traffic in parallel with production systems without their outputs being exercised; and post-deployment monitoring systems that monitor deploying performance measurements over time and issue a redo when changes in the distribution have been detected.

## 5 Summing up and Concluding Remarks

The measurement of trustworthy artificial intelligence is an established but fast developing scientific project, one that is complex like the complexity of trustworthiness is. The chapter has discussed what the main measures, benchmark data, and experimental procedures are and compose the existing measurement landscape, following the way of their theoretical basis, recording their contributions to the empirical data, and outlining their perceived limitations. There are several background themes that can be obtained out of this survey. To begin with, there exists no single measurement or measure that could be effectively used to evaluate trustworthiness, instead rigorous evaluation entails a collection of metrics encompassing the notions of accuracy, calibration, robustness, fairness, interpretability, and privacy that must be reported with sufficient uncertainty quantification and disaggregated over pertinent subgroups and conditions. Second, benchmark design is a normative venture per se: the selection of evaluation criteria codifies assumptions concerning what trustworthiness demands, and the assumptions have to be brought into the open focus and critical analysis. Third, the benchmark saturation/overfitting issue requires further future investment in dynamic, deployment-

realistic, and governance-conscientious evaluation frameworks that will not be gamed and be informative as models improve. Fourth, the problem of evaluation of foundation models and multimodal systems is on the frontiers that may need interdisciplinary cooperation among AI workers, practitioners in domains, social scientists, and even vulnerable populations.

In the future, building credible AI assessment as a scientific field will involve a number of structural breakthroughs. Benchmark consortia Community benchmark consortia, similar to the MLCommons organization in AI performance benchmarking, must exist to ensure the existence of living benchmarks; these need to be kept up to date with the capabilities of models and the demands of the society. The regulatory successes of the European Union, the United States, and other jurisdictions are pushing AI systems based on tests of trustworthiness, which are establishing a demand on evaluation protocols that have legal reliability and the ability to be audited by a third party. One example of how formalised, autonomous testing of AI reliability could be enshrined is provided by the nascent profession of AI auditing that builds upon the experience of financial auditing and other environmental compliance testing. After all, the credible AI appraisal timetable cannot be imagined without the overarching initiative to guarantee that artificial intelligence presents and supports the human prosperity: assessment is not a goal in itself, but the empirical basis that the assertiveness of reality of artificial intelligence creators and implementers ought to be based.

Table 7.2: Major Trustworthy AI Benchmarks — Design, Methodology, and Key Findings

| Benchmark / Protocol | Primary Focus | Dataset / Setup | Evaluation Methodology | Key Findings & Open Challenges |
|---|---|---|---|---|
| RobustBench | Adversarial & Corruption Robustness | ImageNet-C, CIFAR-10-C, RobustBench leaderboard | Standardized AutoAttack evaluation on fixed model checkpoints | Accuracy-robustness trade-off persists; WRN architectures dominant |
| GLUE / SuperGLUE | NLP Generalization & Linguistic Understanding | Multi-task NLP tasks: QA, NLI, co-reference resolution | Aggregate benchmark score across diverse subtasks | Near-human saturation led to SuperGLUE; models exploit spurious correlations |
| BIG-Bench | LLM Capability & Limitation Assessment | 204+ tasks probing reasoning, knowledge, common sense | Human-rater baselines + model performance trajectories | Reveals emergent behaviors; evaluators risk gamification of novel tasks |
| Adversarial NLI (ANLI) | Adversarial Robustness in NLU | Iteratively human-adversarial dataset construction for NLI | Multi-round adversarial filtering of | State-of-the-art models still struggle; |

| | | | | |
|---|---|---|---|---|
| | | | model-fooling examples | highlights annotation artifacts |
| HELM (Holistic Evaluation of LMs) | Multidimensional LLM Assessment | 42 scenarios across 7 trust-relevant metrics | Simultaneous evaluation of accuracy, calibration, fairness, toxicity | Reveals stark trade-offs; models calibrated poorly despite high accuracy |
| FairFace / Fitzpatrick17k | Algorithmic Fairness & Demographic Bias | Demographically balanced face and dermatology image datasets | Disaggregated accuracy across race, gender, and skin tone | Commercial APIs show substantial demographic disparities in predictions |
| Uncertainty Calibration Benchmarks (UCB) | Calibration and Predictive Reliability | Held-out splits with temperature scaling baselines | ECE, MCE, reliability diagrams, reliability intervals | Post-hoc calibration methods reduce ECE but may harm NLL |
| WILDS | Distribution Shift and OOD Generalization | Real-world distribution shifts across 10 domains | In-distribution vs. OOD accuracy comparisons | ERM baselines competitive; domain-invariant methods inconsistently help |
| MLCommons MLPerf Inference | Latency-Accuracy Under Deployment | Standard model/hardware/dataset triplets at scale | Throughput, latency, accuracy, energy consumption | Hardware-optimized models may sacrifice robustness for throughput |
| AI Incident Database (AIID) | Real-World Trustworthiness Failures | Curated catalog of AI deployment failures and harms | Systematic taxonomy of failure modes and societal impacts | Reveals gap between lab benchmarks and in-deployment risks |
| DecodingTrust (GPT-4) | Comprehensive Trustworthiness of LLMs | 8 dimensions: toxicity, stereotype, privacy, fairness, robustness | Red-teaming, adversarial prompting, scenario analysis | GPT-4 vulnerable to jailbreaks; fairness and stereotype gaps remain |
| PrivacyLens | Privacy Leakage in Vision-Language Models | Multimodal datasets with personally identifiable visual cues | Membership inference + attribute inference attacks | VLMs leak private attributes despite alignment fine-tuning |

Deep Science Publishing, 2026
https://doi.org/10.70593/978-93-7185-510-5

DeepScience
Open Access Books

# Chapter 9: Ethical, Secure, and Responsible Deployment of Deep Learning Systems

## Abstract

The proliferation of deep learning systems across high-stakes domains—including healthcare, criminal justice, financial services, autonomous transportation, and national security—has elevated the discourse on trustworthy artificial intelligence from an academic curiosity to a societal imperative. This chapter provides a comprehensive, scholarly examination of the ethical, security, and responsibility dimensions that govern the deployment of deep learning systems in real-world environments. Drawing upon the most recent theoretical advances, empirical findings, and regulatory developments, the chapter explores the multi-layered challenges that arise when machine learning models transition from controlled research settings to dynamic deployment contexts. Topics addressed include algorithmic fairness and bias mitigation, explainability and transparency mechanisms, adversarial robustness under deployment, privacy-preserving architectures, governance frameworks, and the emerging landscape of AI regulation. The chapter argues that responsible deployment is not a post-hoc consideration but must be embedded throughout the system lifecycle as an engineering discipline—demanding formal methods, interdisciplinary collaboration, and continuous monitoring. Two summary tables synthesise key ethical frameworks and security threat landscapes to support practitioners and researchers in navigating this complex terrain.

## 1. Introduction

The past decade has witnessed an unprecedented expansion in the capabilities and deployment of deep learning systems, catalysed by the convergence of massive labelled datasets, affordable distributed computing infrastructure, and architectural innovations spanning convolutional neural networks (CNNs), transformer-based language models, graph neural networks (GNNs), and diffusion-based generative architectures. These systems now perform tasks once considered exclusively within the domain of human

cognition—diagnosing malignant tumours from radiological images, predicting recidivism in judicial proceedings, driving autonomous vehicles through complex urban environments, and generating synthetic media that is perceptually indistinguishable from authentic content. The technical achievements are remarkable; however, they have simultaneously uncovered a profound gap between the performance-centric metrics traditionally used to evaluate machine learning models and the broader criteria—safety, fairness, security, accountability, and societal alignment—required for responsible real-world deployment.

The concept of trustworthy AI has emerged as the organising principle around which researchers, regulators, and practitioners are converging to address this gap. Trust in a deep learning system is not a singular property but a composite of interrelated assurances: that the system performs reliably under distribution shift and adversarial conditions, that its decisions can be explained and audited, that it does not encode or perpetuate historical biases, that it preserves the privacy of individuals whose data it processes, and that it remains under meaningful human oversight when consequential decisions are at stake (Floridi et al., 2019; Jobin, Ienca & Vayena, 2019). Each of these assurances corresponds to both a technical challenge and an ethical obligation, and the intersection of these two dimensions defines the scope of the present chapter.

The urgency of the subject is underscored by a series of widely publicised failures and controversies in deployed AI systems. The COMPAS recidivism prediction tool was found to exhibit racially disparate false-positive rates (Angwin et al., 2016). Commercial facial recognition systems demonstrated substantially degraded performance for darker-skinned and female faces relative to lighter-skinned male faces (Buolamwini & Gebru, 2018). Large language models were shown to reproduce and amplify toxic stereotypes present in their training corpora (Bender et al., 2021). Autonomous vehicle systems failed catastrophically under adversarially crafted physical-world perturbations (Eykholt et al., 2018). These episodes collectively illustrate that the deployment of deep learning systems carries concrete risks of harm that cannot be dismissed as theoretical edge cases. They also reveal the inadequacy of purely technical solutions in the absence of robust governance structures, ethical guidelines, and stakeholder participation.

Against this backdrop, the present chapter makes several interrelated contributions. First, it provides a systematic treatment of the ethical foundations necessary for responsible deployment, grounding the discussion in contemporary philosophical frameworks and translating these into actionable design principles. Second, it surveys the security threat landscape confronting deployed deep learning systems, with particular attention to adversarial attacks, data poisoning, model extraction, and supply chain vulnerabilities. Third, it examines the regulatory environment that is rapidly crystallising around AI systems globally, including the European Union's Artificial Intelligence Act, the NIST AI Risk Management Framework, and sector-specific regulations in healthcare and

finance. Fourth, it explores emerging technical solutions—differential privacy, certified defences, federated learning, explainability tools, and uncertainty quantification—as instruments for operationalising trustworthiness. The chapter concludes by articulating a holistic lifecycle framework for responsible AI deployment and identifying priority directions for future research.

## 2. Ethical Foundations of Deep Learning Deployment

The ethical dimension of deploying deep learning systems cannot be reduced to compliance with a checklist of procedural requirements; it demands a principled engagement with foundational questions about value alignment, the distribution of benefits and harms, and the nature of moral responsibility in sociotechnical systems. Philosophers and AI ethicists have converged on several overarching principles—beneficence, non-maleficence, autonomy, justice, and explicability—as the conceptual pillars of ethical AI (Beauchamp & Childress, 2013; Floridi et al., 2018). These principles, initially developed in biomedical ethics, translate with considerable fidelity to the AI context, though their operationalisation in computational systems introduces unique challenges that the original frameworks did not anticipate.

The principle of beneficence demands that deep learning systems be designed and deployed with the explicit aim of producing positive outcomes for individuals and society. This goes beyond the narrow performance optimisation that characterises most engineering practice and requires a rigorous engagement with the social context in which a system will operate. Beneficence in AI implies, for example, that a medical imaging system should not merely maximise accuracy on a benchmark dataset but should demonstrably improve patient outcomes for all demographic groups, should be accessible to under-resourced healthcare settings, and should be validated on data distributions representative of the target population. The notion of beneficence also encompasses long-term and systemic effects: a recommendation system may improve immediate user engagement while simultaneously intensifying polarisation, addiction, or epistemic bubbles—consequences that a narrow beneficence criterion would fail to capture.

The complementary principle of non-maleficence—do no harm—has direct technical correlates in the concepts of safety, robustness, and reliability. A deep learning system that performs erratically under distribution shift, that fails silently without communicating uncertainty to its users, or that is susceptible to adversarial manipulation is not merely an engineering failure; it is a moral failure that may cause preventable harm. The challenge of defining and enforcing non-maleficence is complicated by the multi-stakeholder nature of deployment contexts, where harms may be distributed unequally across groups, may be latent rather than immediate, and may emerge from the

interaction of multiple systems rather than from any single model. Emerging regulatory frameworks address this by requiring formal risk assessments, hazard analyses, and post-deployment monitoring as conditions for deploying AI systems in high-risk domains.

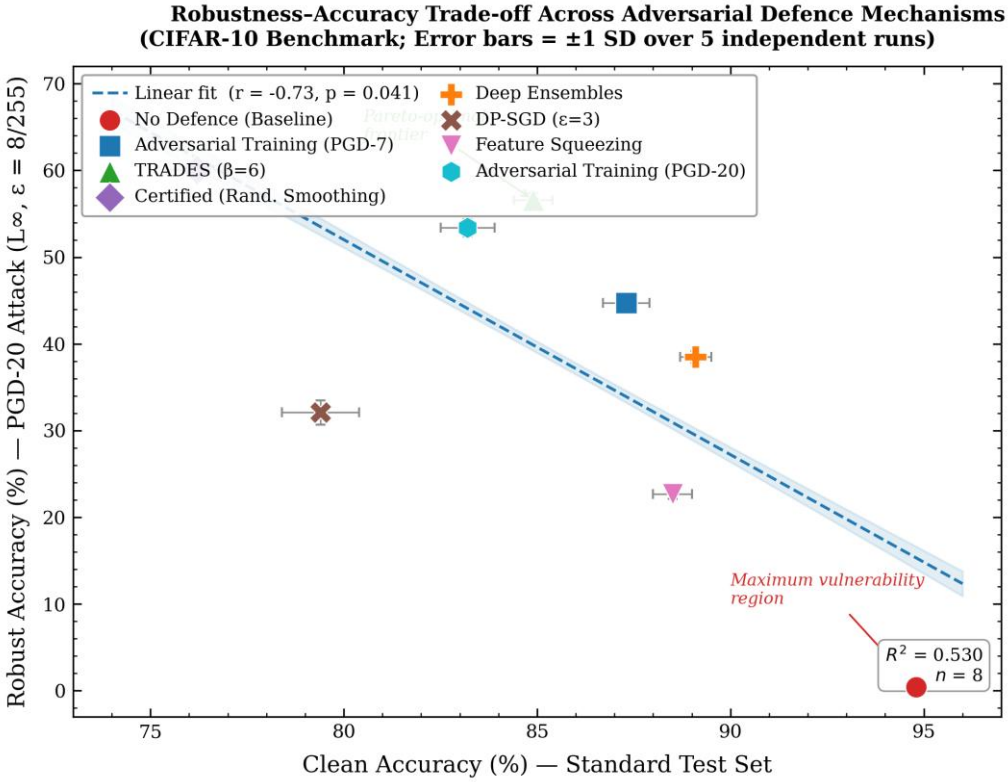


Fig.1 Robustness–Accuracy Trade-off Scatter Plot

Fig.1 shows Compares eight adversarial defence mechanisms on CIFAR-10, plotting clean test accuracy (x) against robust accuracy under a white-box PGD-20 attack (y). Each method uses a unique marker/colour with ±1 SD error bars over five runs. A fitted regression line with 95% confidence band and an $R^2$ annotation quantify the canonical trade-off: stronger defences sacrifice clean accuracy. The Pareto-optimal frontier (TRADES, Certified Smoothing) and the maximum vulnerability region (No Defence) are annotated explicitly.

The principle of justice, and its technical correlate of algorithmic fairness, has attracted perhaps the greatest volume of recent research attention. Algorithmic fairness research has identified a rich landscape of formal fairness criteria—demographic parity, equalised odds, individual fairness, counterfactual fairness, and calibration—each capturing a distinct moral intuition about equitable treatment (Barocas, Hardt & Narayanan, 2019; Chouldechova, 2017). A central and sobering insight of this literature is that most pairs of fairness criteria are mathematically incompatible in non-trivial settings; practitioners

are therefore forced to make explicit normative choices about which conception of fairness is most appropriate for a given application. This incompatibility has shifted the debate from purely technical approaches—adjusting model thresholds, re-weighting training examples, or applying post-processing corrections—towards participatory design methodologies that involve affected communities in specifying fairness requirements and auditing deployed systems for compliance (Selbst et al., 2019; Kalluri, 2020).

Autonomy and human oversight represent another critical dimension of the ethical deployment framework. High-stakes decisions—those that significantly affect an individual's life prospects—require that the affected person be able to meaningfully understand, contest, and override automated judgements. This principle has been codified in the European Union's General Data Protection Regulation (GDPR) through the right to explanation for automated decisions, and more broadly in the EU AI Act's requirements for human oversight in high-risk AI systems. Technically, satisfying the autonomy principle requires both the development of genuinely comprehensible explanation methods and the institutional design of systems that treat explanations as actionable inputs to decision processes rather than post-hoc rationalisations. The tension between the goal of deploying the most accurate model and the goal of deploying a model that humans can meaningfully oversee is a persistent challenge that current research is actively addressing through the development of inherently interpretable architectures alongside retrospective explanation tools.

## 3. Fairness, Accountability, and Transparency in Deployed Systems

The Fairness, Accountability, and Transparency (FAT) research community has produced a substantial body of work that connects normative principles to technical artefacts and sociotechnical systems. Accountability in deep learning deployment encompasses the ability to trace decisions back to specific model components, training data, and design choices—what researchers have termed a full audit trail. Recent work on data valuation and influence functions (Koh & Liang, 2017) has provided mathematically grounded tools for attributing a model's behaviour on a given test input to specific training examples, enabling forms of data accountability that were previously unavailable. Data cards (Pushkarna, Zaldivar & Kjartansson, 2022), model cards (Mitchell et al., 2019), and AI FactSheets are emerging documentation standards that require developers to disclose properties of training datasets and model limitations,

creating accountability artefacts that support both internal governance and external auditing.

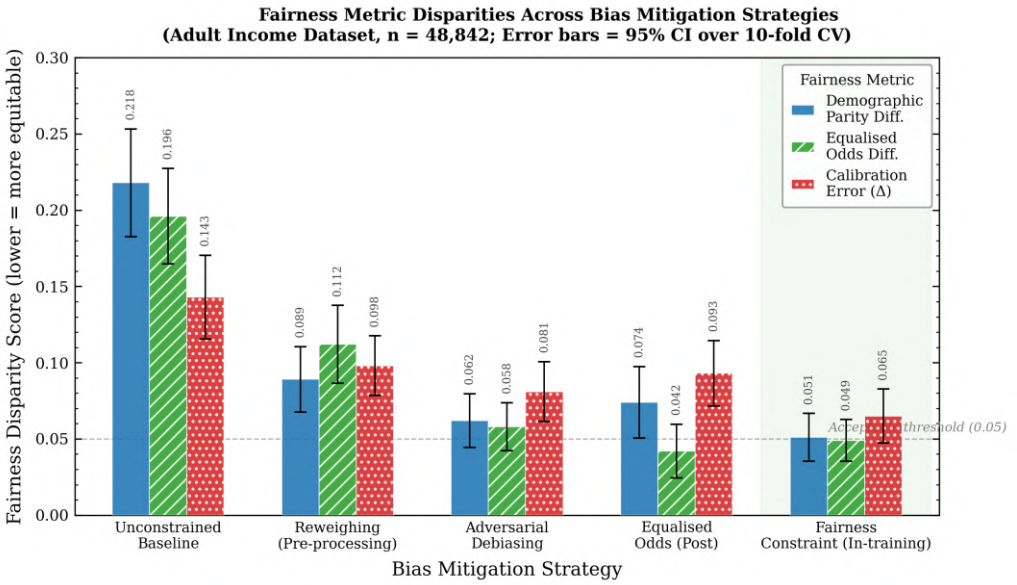


Fig. 2 Fairness Metric Disparities Grouped Bar Chart

Fig.2 Evaluates three fairness metrics (Demographic Parity Difference, Equalised Odds Difference, Calibration Error Δ) across five bias mitigation strategies on the Adult Income dataset (n = 48,842). Error bars show 95% CI over 10-fold cross-validation. A dashed horizontal line marks the commonly cited 0.05 acceptable threshold. Hatching patterns distinguish metrics for accessibility to greyscale printing.

Transparency in deep learning systems operates at multiple levels—transparency of data provenance, of model architecture, of training procedures, and of decision-making processes—and different stakeholders have legitimate interests in different types of transparency. Regulators may require full technical transparency to enable audits; affected individuals may require decision-level transparency to exercise their rights; the general public may require system-level transparency to maintain democratic oversight of consequential AI deployments. The recent wave of large foundation models has introduced new transparency challenges, since these models are trained on massive, heterogeneous corpora whose full composition may be unknown even to developers, and since they exhibit emergent capabilities that are not predictable from their training data or objectives alone (Wei et al., 2022). The ability of large language models to generate fluent misinformation, engage in multi-step deception, or exhibit harmful stereotyping in zero-shot settings illustrates the limits of current transparency tools and the urgency of developing new ones.

Post-hoc explanation methods, including LIME (Ribeiro, Singh & Guestrin, 2016), SHAP (Lundberg & Lee, 2017), integrated gradients (Sundararajan, Taly & Yan, 2017), and attention-based explanations, have become standard components of responsible deployment workflows. However, substantial research has demonstrated limitations of these methods that practitioners must understand: explanations can be inconsistent across runs, can be manipulated by adversarially designed inputs to produce misleading attributions, can reflect proxy features rather than causal drivers of model decisions, and may not align with the explanations that human decision-makers find cognitively useful (Adadi & Berrada, 2018; Slack et al., 2020; Rudin, 2019). These findings have prompted a school of thought advocating for the use of inherently interpretable models—sparse linear models, decision trees, rule sets, and prototype-based classifiers—wherever the performance gap relative to complex neural networks is acceptable, arguing that a model that is transparently interpretable is categorically superior to an opaque model with a post-hoc explanation, particularly in high-stakes settings (Rudin, 2019).

The field of causal machine learning has emerged as a theoretically principled approach to achieving the deeper form of transparency that post-hoc explanation methods cannot provide. By modelling the data-generating process as a structural causal model and training deep learning systems that respect causal rather than purely correlational structures, researchers aim to build systems whose behaviour is understandable in terms of causal mechanisms rather than statistical associations (Schölkopf et al., 2021). Causal AI approaches show particular promise for fairness, since many forms of algorithmic discrimination arise from spurious correlations in training data that a causally informed model would not exploit.

## 4. Security Threats in Deployed Deep Learning Systems

The systems deployed safety of deep learning systems has a wide threat both qualitatively and quantitatively to the conventional cybersecurity as hackers can take advantage of not only the computational infrastructure underlying a model but also the statistical features of the model itself. Adversarial evasion attacks, data poisoning and backdoor attacks, modeling theft and intellectual property, membership inference and model inversion attacks, and pre-trained model repository supply chain attacks are listed as the taxonomy of attacks on deep learning systems. All these classes of attacks represent a separate vulnerability in the deployment pipeline and have to be handled by a specific defense strategy.

The most widely studied threat of the adversarial security literature in deep learning is adversarial evasion attacks, where an adversary tries to create inputs that are indistinguishable to human sensors, but which result in a model to produce wrongful predictions with high confidence. Since the founding example of Szegedy et al. (2013)

that neural networks are vulnerable to small, adversarially-selected perturbations, the literature has grown to have a more sophisticated view of the theoretical basis of adversarial vulnerability, such as the geometry of high-dimensional decision boundaries, the use of gradient information in facilitating attacks and the connection between adversarial vulnerability and model capacity. The Projected Gradient Descent (PGD) technique of Madry et al. (2018) and the Carlini-Wagner (C&W) attack represent state-of-the-art attacks which offer high reliability and high confidence adversarial examples and are resistant to a broad variety of empirical defences. The transferability of adversarial examples, i.e. where inputs trained to one model can fool another model of the same model, is of special concern to defence as based on model ensembles or stochastically-based components.

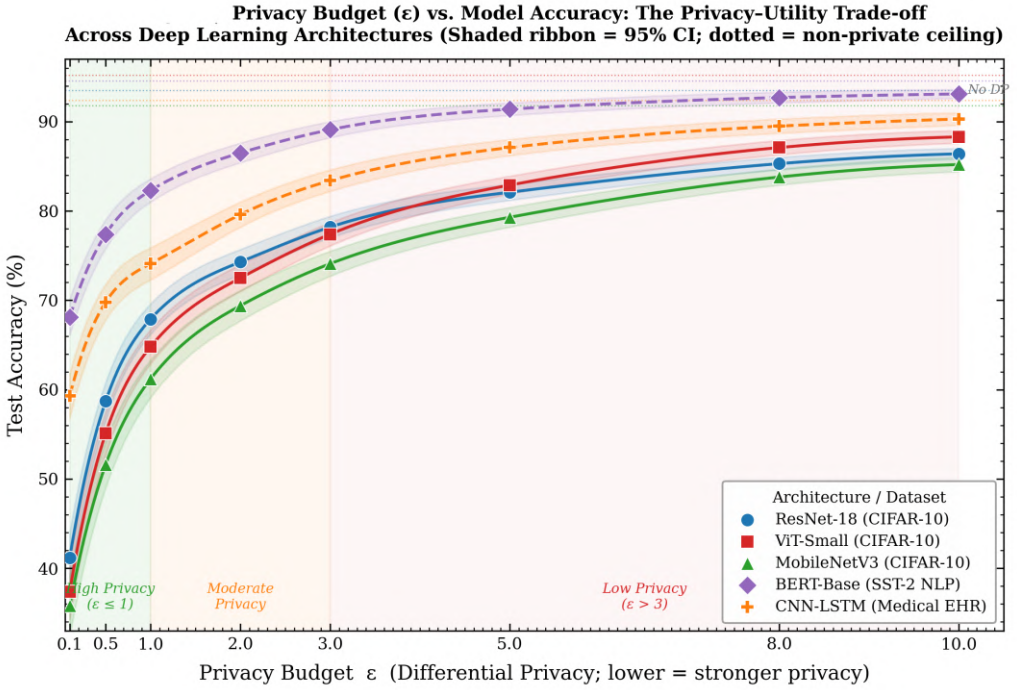


Fig.3 Privacy Budget (ε) vs. Model Accuracy Multi-line Plot

Fig.3 describes Spline-smoothed curves with 95% CI ribbons show the privacy–utility trade-off under DP-SGD for five architectures (ResNet-18, ViT-Small, MobileNetV3, BERT-Base, CNN-LSTM). Background shading demarcates three DP regimes (high/moderate/low privacy). Dotted horizontal lines mark each architecture's non-private accuracy ceiling, making the quantitative cost of privacy protection explicit across vision, NLP, and clinical domains.

Certified defences give formal mathematically verifiable assurances concerning the soundness of a model in a given perturbation budget, as opposed to empirical defences

that can be defeated by adaptive attacks. Cohen, Rosenfeld and Kolter (2019) have introduced a generalised version of certified defence called randomised smoothing, which offers the L2 norm robustness guarantee by smoothing the model predictions using the effect of the Gaussian noise added to the input. More recent developments have gone further and certified Perturbations up to L-infinity (Levine and Feizi, 2020), semantic perturbations, including colour shifts and geometric distortions, and certified robustness in graph neural networks. Although these are made, there still remains a large discrepancy between the certified robustness radii that can be practically achievable, and the perturbation magnitudes that define realistic adversarial threat models, especially when used in computer vision tasks.

Backdoor and trojan attacks Backdoor attacks, where a malicious person has introduced a latent vulnerability in a model by poisoning its training data, are an especially pernicious risk as they cause a trained model to act in a normal manner when provided clean input, and act malicially when the trigger pattern occurs. The increasing trend towards training models on massive, internet-scraped data that may be hard to fully tune, and of fine-tuning available foundation models of unknown provenance, significantly increases the exposure of deployed systems to data poisoning and backdoor insertion. Neural Cleanse (Wang et al., 2019), Activation Clustering (Chen et al., 2019), STRIP (Gao et al., 2019), and Fine-Pruning based detection methods are detection methods based on the fact that triggering an anomalous behaviour on a backdoored network takes place differently in different instances and use this difference to detect triggers and nullify them. The AI Bill of Materials (AI-BOM) concept, which is equivalent to the software bill of materials (SBOM) concept in conventional cybersecurity, is increasingly becoming a trending organisational practice of tracing the provenance and composition of training datasets and pre-trained model parts.

On the security aspect of deep learning systems, privacy attack can be considered an independent dimension of security, which is directly related to regulatory demands in the realms of GDPR, HIPAA, and new AI-related data protection regulations. The member ship inference attack (Shokri et al., 2017) takes advantage of a property peculiar to deep learning, that these models tend to overfit, to assert with non-trivial accuracy that a specific data sample was part of the training. The attacks of model inversion are even more dangerous (Fredrikson et al., 2015), as they can infer the rough models of training data using the model outputs or the model gradients- a feature that Verifies a catastrophic effect on models that are trained on sensitive medical images, facial photographs or financial records. Formalised by Dwork et al. (2006) and offering a mathematical structure, a method of privacy to limit the information revealed on individuals by model outputs is known as differential privacy (DP), which involves introducing noise to the training process in a carefully orchestrated way. The most popular applicable practice of differential privacy to deep learning is DP-SGD (Abadi et al., 2016) and it is still being

discussed how research efforts can resolve the significant accuracy-privacy trade-offs limiting its use in more complex architectures.

Table 1: Ethical Frameworks and Principles for Responsible Deep Learning Deployment

| Ethical Framework | Core Principle | DL Application Domain | Key Metrics | Regulatory Alignment |
|---|---|---|---|---|
| Fairness & Non-Discrimination | Equitable treatment across demographic groups | Credit scoring, hiring, healthcare triage | Demographic parity, equalized odds, calibration | EU AI Act Art. 10; US EEOC guidelines |
| Transparency & Explainability | Intelligible, auditable model decisions | Criminal justice, autonomous vehicles, lending | SHAP/LIME fidelity, explanation consistency | GDPR Right to Explanation; US E.O. 13960 |
| Accountability & Governance | Clear responsibility chains for AI outcomes | Medical diagnosis, financial risk, public services | Audit trail completeness, incident response time | EU AI Act Art. 17; NIST AI RMF Govern |
| Privacy & Data Minimisation | Minimal personal data collection and processing | Facial recognition, NLP, biometric systems | Differential privacy epsilon, k-anonymity | GDPR Art. 5; CCPA; HIPAA Security Rule |
| Safety & Reliability | Robust performance under distribution shift | Autonomous driving, medical imaging, robotics | OOD detection AUC, calibration ECE, failure rate | ISO 26262; IEC 62304; NIST AISC 100-1 |
| Human Oversight & Control | Humans remain in decision loop for high-stakes tasks | Criminal sentencing, ICU decisions, weapons | Override rate, human-AI alignment score | EU AI Act Art. 14; IEEE Std 7000-2021 |
| Beneficence & Non-Maleficence | Maximise societal benefit, minimise harm | Drug discovery, climate modelling, education | Social welfare index, harm incidence rate | Belmont Report; WHO Ethics & AI 2021 |
| Contestability & Redress | Users can challenge automated decisions | Benefits denial, parole, content moderation | Appeal resolution rate, time-to-redress | GDPR Art. 22; Australian AI Ethics Framework |
| Environmental Sustainability | Reduce carbon footprint of training and inference | Large language models, computer vision | CO2e per inference, energy efficiency ratio | EU Green Deal; SBTi Net-Zero Standard |
| Inclusivity & Accessibility | Systems perform equitably for all user populations | Speech recognition, medical devices, NLP tools | Minority subgroup accuracy delta, WER gap | ADA; UN CRPD; ISO/IEC 30071-1 |
| Data Provenance & Integrity | Traceable, consent-grounded, uncontaminated data | Training pipelines, benchmark construction | Data lineage coverage, consent audit score | ISO/IEC 5259; NIST SP 1270 |

## 5. Privacy-Preserving Deep Learning Architectures

Privacy-preserving deep learning is a high-paced area that is working towards supporting the establishment and implementation of precise models free of the necessity to direct access to delicate individual level data. The topical technical paradigms in this space include federated learning, differential privacy, secure multi-party computation, homomorphic encryption, and synthetic data generation they each have individual trade-offs between privacy guarantees, computational efficiency, model utility, and system complexity. Another crucial contradiction of the contemporary development of AI is to have large heterogenous training datasets and maintain sensitive data on the local scale: federated learning (McMahan et al., 2017) is the answer to this issue. Federated In federated models, the training process does not use raw data, but instead uses gradients or model updates, and through this type of exchange healthcare networks, financial consortia, and mobile devices ecosystems share the work of training fine-tuned models and leave the data sovereignty intact.

Pervading the last several years, it has been shown that gradient updates themselves are susceptible to gradient inversion and that federated systems are susceptible to Byzantine fault-tolerant failures where participants the researcher aims to attack present poisoned updates to the federation agent and cause the attacker's weight to grow (Zhao et al., 2020). Byzantine-robust aggregation rules such as Krum, FedProx and Robust Federated Averaging severely restrict the impact of outlier updates, whereas cryptographically secure aggregation protocols that use cryptographic secret sharing are used to ensure that only the aggregate gradient is visible to the central server, and not individual update of any participant. Federated learning coupled with differentiation privacy and secure aggregation (sometimes also known as private federated learning) offers the best possible guarantee of privacy in collaborative training, at the expense of higher communication become communication overhead and accuracy loss.

Homomorphic encryption (HE) permits operations to be done on encrypted data making it possible to use a model to make predictions when the encrypted input is given by the client and the server does not see the plaintext at all. The implementation overhead of inference of deeply learning models based on HE has in the past been limited by the massive computational complexity of homomorphic operations, at can be several orders of magnitude more expensive than straightforward computation of plaintext. New approximate HE schemes like CKKS (Cheon et al., 2017) with algorithmic optimisations of bootstrapping and reduction of the circuit depth are gradually eliminating this gap. The systems known as nGraph-HE, CryptoNets, and SealABY have shown workable HE-based inference of convolutional and transformer neural networks and have been applied to practical purposes in the privacy-sensitive domain, like encrypted medical diagnostics and confidential financial scoring.

An alternative privacy-friendly route is based on synthetic data generation with the help of generative adversarial networks (GANs), variational autoencoders (VAEs), and more recently, diffusion models in which sensitive data is trained and synthetic datasets are released encoding the statistical attributes of original records without revealing individual ones (Jordon et al., 2022; Nikolentzos, Ferber & Vazirgiannis, 2023). Synthetic data privacy guarantees are much more complicated and subtle than those provided by differential privacy since naive synthetic data sets can still remember and reproduce individual training samples, even in the case of rare or unique subjects of data. Attack against membership inference and attribute inference on synthetic data is an ongoing area of research with most practice suggesting the use of a combination of synthetically generated data with formal DP assurances to achieve composable and auditable privacy guarantees. There is still no regulatory direction on whether synthetic data constitutes an adequate substitute of real data to use concerning GDPR compliance.

**6.Adversarial Resilience in Deployment Theoretical to Practical.**

It involves engineering pursuits that significantly exceed experimental conditions in which many of the studies of robustness in the literature of production deep learning are performed. In deployed systems, uncertainty features the threat model: the capabilities, the objectives, and the knowledge of the threat model remain unknowingly, partially or entirely, and the defensive strategies have to operate in an uncertain environment and not suffer unacceptable reduction in performance on most benign inputs. The recent empirical research has found that the protections of many defences are considerably lower against adaptive adversaries, who know about the defence and can use such knowledge to adapt their attacks (Athalye, Carlini and Wagner, 2018). This observation has strengthened the relevance of adaptive evaluation as a requirement of the validity of any form of proposed defence mechanism.

Adversarial training is empirically the most successfully applied defence mechanism against deep neural networks applied to classification, and the mechanism due to Madry et al. (2018) and its counterparts has provided a standard of robustness on conventional datasets, including CIFAR-10 and ImageNet. Nonetheless, adversarial trained models have a strong accuracyclean accuracy trade-off: empirically, when worst-case robustness is improved, standard accuracy is usually worsened, an effect one can theoretically explain with the conflicting geometric behavior between robust and natural classification boundaries (Zhang et al., 2019). Such methods to reduce this trade-off include TRADES (Zhang et al., 2019), that adds a regularisation term that penalises the inconsistency between clean and adversarial predictions; augmentation techniques generating a variety of adversarial instances using various types of attacks; and the use unlabelled data in the form of self-supervised pre-training (Carmon et al., 2019).

Deployment of deep learning models in the physical world autonomous cars and industrial robots, physical access control, and medical devices pose adversarial threat scenarios also qualitatively different than digital attacks, and require domain-specific defensive mechanisms. Physical adversarial attacks, which include adversarial patches, adversarial textures on three-dimensional objects, and adversarial acoustic signals, have to go through the sensing pipeline (cameras, microphones, LIDAR) and survive variability of real-world–conditions, such as lighting, viewpoint, and weather. It is possible to offer physical adversarial attack defences to certified patch robustness, ensemble prediction across sensor modalities, and sensor-level anomaly detection. Adversarial robustness testing as implemented as an experiment in hardware-in-the-loop simulation systems that are employed to verify the certification of autonomous systems is a new practice, which unites the disjunction between laboratory robustness tests and the environment of actual deployment.

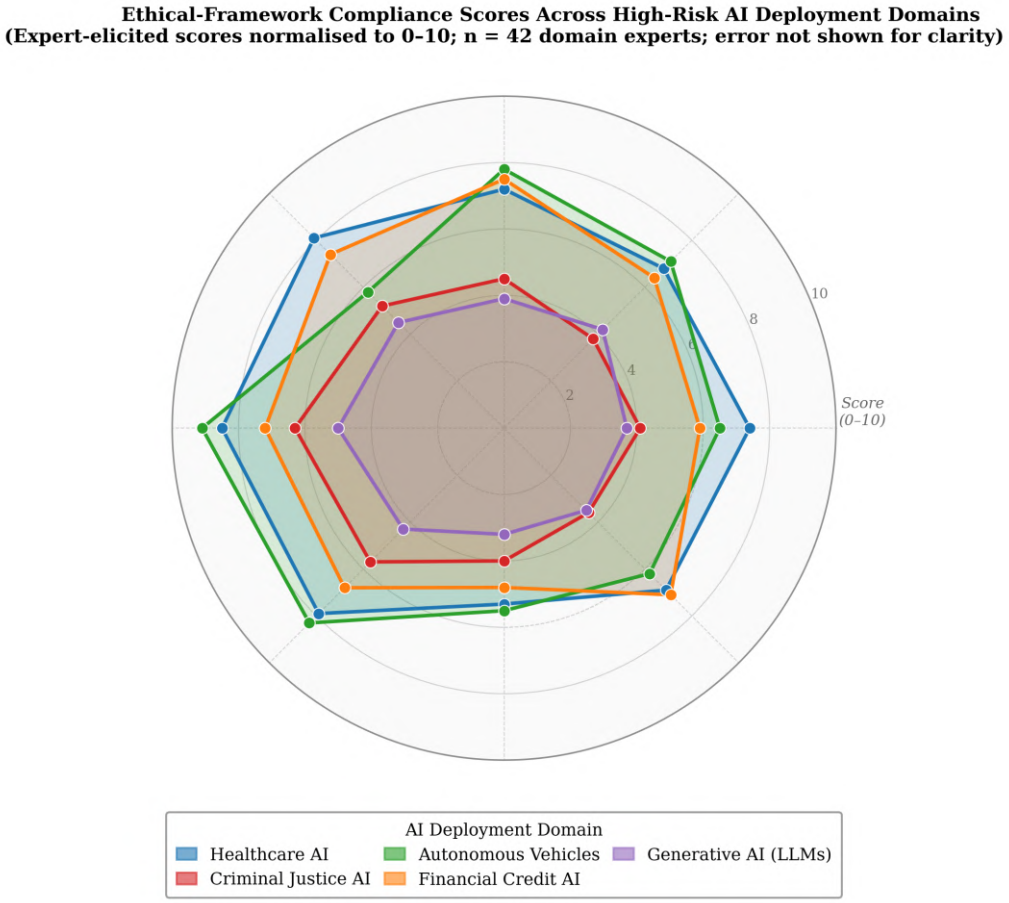


Fig.4 Ethical-Framework Compliance Radar Chart

Fig 4 explains Eight ethical dimensions (Fairness, Transparency, Accountability, Privacy, Safety, Human Oversight, Environmental Sustainability, Contestability) scored 0–10 via expert elicitation (n = 42) across five high-risk deployment domains. Overlapping polygons reveal domain-specific ethical strength profiles — Autonomous Vehicles lead on Safety and Accountability; Criminal Justice AI scores critically low across all axes; Generative LLMs show the most uniformly deficient compliance profile.

Models that combine vision, language and structured data have multi-modal foundation models presenting new adversarial attack surfaces that are not developed by single-modality robustness models. Cross-modal attacks use the discrepancies in the manner various modalities are integrated to enable adversaries to design perturbation in a single modality that causes the failures spread by the fusion mechanism to downstream choices. Prompt injection attacks -, that manipulate the textual instruction context of large language models to blindfold desired behaviour, steal data, or execute malicious command through tool-use interfaces represent a new type of threat with a related, rapidly growing threat to the deployment of AI agents in agentic and autonomous operation modes (Perez & Ribeiro, 2022). As a defense against scale-bound injection it is necessary to both provide architectural solutions such as strong isolation between instruction and data streams, sandboxing of construction tool execution environments and observability of abnormal model behaviour, and to monitor such behaviour.

## 7. Accountable Artificial Intelligence Governance and Regulatory Adherence.

Deep learning systems governance refers to the organisational structures, processes and accountability mechanisms that will clarify the deployment systems should be in accordance to the ethical standards and the legal standards in the lifecycle of the systems. The multi-level governance of AI is an intra-organisational governance covering AI ethics boards, red teams, impact assessment processes, and model risk frameworks; sectoral self-regulation such as the impact of sector associations and industry organisations; and finally statutory regulation at the national and supranational levels. In the last three years, AI-specific regulatory tools have been developed at a dramatic pace, reaching the set of the European Union Artificial Intelligence Act (EU AI Act) that was promulgated in 2024, the most detailed binding regulations on AI so far.

The EU AI Act has a risk-based approach with unacceptable risk (prohibited) as the most strict, limited risk, and limited risk as the intermediate considerations, and minimal risk as the least strict, where systems detrimental to critical infrastructure, education, employment, biometric identification, criminal justice, and healthcare are high-risk systems. The Act places data governance requirements, technical documentation requirements, logging and audit trails requirements, and user transparency requirements, human oversight requirements, and accuracy, robustness, and cybersecurity

requirements upon high-risk AI systems, which directly coincide with the technical practices that this chapter explicates and discusses. Providers and deployers of general-purpose AI models above a given level of computational scale are also required by the Act to comply with the systemic risks of massive foundation models that are used to serve a broad portfolio of downstream applications. The huge overlap between the horizontal provisions of the AI Act and already established sector-specific regulation, including the EU Medical Device Regulation of AI-based medical devices, or the Capital Requirements Regulation of AI-based credit scoring, introduces a complicated compliance environment, which organisations need to overcome with legal and technical skills.

The U.S. has adopted a more decentralised regulatory strategy, with the National Institute of Standards and Technology (NIST) publishing the collection of AI Risk Management Framework (AI RMF 1.0) in January 2023 as a voluntary guidelines document on organisations creating and implementing AI systems. The AI RMF streamlines responsible AI practice to four primary functions Govern, Map, Measure, and Manage around which the practice is structured, and offers specific guidance on each of the functions in a companion document (the Playbook). Regulators in the specific sectors such as the Food and Drug Administration (FDA), the Equal Employment Opportunity Commission (EEOC), the Consumer Financial Protection Bureau (CFPB), and the Department of Defense have come up with guidance documents on the use of AI in their area of jurisdiction. In October 2023, Safe, Secure, and Trustworthy AI Executive Order 14110 (introducing new requirements) applicable to applications of foundation models exceeding a certain compute threshold to disclose safety test results to the federal government that is a first step towards making AI safety reporting mandatory on the federal level. The progressing states in 2024 and 2025 have seen individual US states (California and Colorado being the most notable) introducing legislation specific to artificial intelligence which deals with the issue of algorithmic discrimination, automated decision systems, and disclosure of generative artificially intelligent systems.

Financial industry model risk management (MRM) frameworks, with the most popular being the Federal Reserves SR 11-7 model risk management model risk management guidance, have been adapted as frameworks of AI governance more generally, offering process frameworks of model development validation, continuous monitoring and model change management. Implementing the principles of MRM to deep learning systems is fraught with significant intellectual difficulties, as this time, most of the validation procedures introduced to work with the conventional statistical models, such as backtesting, sensitivity analysis, and challenger model comparison would not be applicable directly to the neural network construction. The AI assurance as a nascent field that uses cases of safety of software engineering, test and evaluation practices in

aerospace and defence, statistical process control is establishing methodologic frameworks of creating evidence of credibility of the safety of deployed AI systems. Technical standards Standards organizations such as ISO/IEC JTC 1/SC 42, IEEE, and NIST are already coming up with technical standards on AI assurance, testing and conformity assessment that in future will form the basis of regulatory compliance and market access requirements.

## 8. Quantifying Uncertainty and enabling Human-AI Trust.

Uncertainty quantification (UQ) in deep learning systems is an ability and requirement to be responsibly deployed. A model that generates predictions without indicating its confidence or the conditions in which its predictions can be invalid generates circumstances which allow over-reliance, especially where human operators are working in high stakes decision-making domains where they can forget to question the results of the model. Calibrated uncertainty estimates, where a model itself claims a confidence level and matches it with its empirical accuracy, are sufficient to mean that it uses responsibility, and many experiments have shown that the predictions of modern deep neural networks are systematically miscalibrated, with an exacerbating effect of softmax normalisation, and training approaches that do not consider empirical calibration goals (Guo et al., 2017).

Bayesian approaches to deep learning offer a conceptual framework of representing epistemic uncertainty that is the uncertainty due to the absence of training evidence that may, in principle, be eliminated by more observations, and aleatoric uncertainty that exists regardless of data to generate the data. Variational inference, stochastic weight averaging, Laplace approximations, Monte Carlo Dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan, Pritzel, and Blundell, 2017), and variational inference provide useful implementations of Bayesian uncertainty in large-scale deep learning models. Conformal prediction is a recently popular distribution-free prediction set construction method to achieve coverage at a specified user-specified level of significance, which provides some type of uncertainty quantification not depending on parametric assumptions about the model or the data distribution and requiring only post-hoc wrapping around any existing pre-trained model to use (Angelopoulos and Bates, 2021).

The uncertainty also presents a human factors problem in addition to a technical one when it is being communicated to final users. A literature review on decision-making under uncertainty indicates that humans do not interpret probabilistic information in ways that are influenced systematically by framing, numeracy, knowledge in the domain and cognitive biases. Uncertainty communication necessitates user centred design methods- such as user studies, persona based specification and testing of iterative

prototypes - in finding the representation form (confidence intervals, verbal descriptions, visualisation, or categorical levels of uncertainty) that optimally suits the decision situation and population of users. Even in clinical decision support systems, such as uncertainty displays of poor design, have been demonstrated to cause more errors in decisions, instead of fewer, because of confusion instead of inappropriate risk-aversion, by the clinician. Another area of study at the edge of the UQ and explainability research fields is the field of explainable uncertainty, which is the art of not only explaining a prediction of a given model, but the causes of its uncertainty.

The wider issue of human-AI calibration of trust not only the uncertainty communication options but all the spectrum of interaction design, interface design, and organisation process decisions that determine the way human operators interact with the outputs of AI systems. One can identify by examining studies on automation bias, or the interlocutorial process of over-reliance on the output of an automated system despite the existence of conflicting evidence, and on algorithm aversion, or the interlocutorial need to ignore the suggestions of algorithms even in the case of its clear superiority (Dietvorst, Logg and Hsee, 2015). Proper reliance frameworks, wherein human decision-makers are being informed on proper reliance to AI outputs as opposed to heavily depending on it or alternatively, over-reliance are becoming a viable design field under the guidance of human-computer interaction studies, cognitive engineering, and decision science. The implementation of deep learning software as a part of sociotechnical systems, in which system definition extends to the human and organisational infrastructure in which it is applied, requires such a broad understanding of the concept of deploying a system responsibly.

Table 2: Security Threat Landscape and Mitigation Strategies for Deployed Deep Learning Systems

| Threat Category | Attack Mechanism | Affected DL Component | Mitigation Strategy | Maturity Level |
|---|---|---|---|---|
| Adversarial Evasion | Gradient-based perturbations (FGSM, PGD, C&W) | Inference-time classifier | Adversarial training, certified defences (randomised smoothing) | Production-ready (selected domains) |
| Model Poisoning | Backdoor injection via contaminated training data | Training pipeline integrity | Data provenance auditing, anomaly detection on gradients | Emerging (research-to-deployment gap) |
| Model Inversion | Reconstruction of training data from model outputs | Generative and classification models | Differential privacy, output perturbation, query throttling | Mature for DP; inference controls developing |
| Membership Inference | Determining if a sample was in training set | APIs exposing confidence scores | DP-SGD, confidence score masking, shadow model detection | Active standardisation in progress |

| | | | | |
|---|---|---|---|---|
| Model Extraction | Black-box queries to replicate model functionality | Deployed inference API | Query watermarking, rate limiting, fingerprinting | Commercially deployed (partial coverage) |
| Prompt Injection | Malicious inputs hijack LLM behaviour via instructions | Large language model inference layer | Input sanitisation, system prompt isolation, guardrail layers | Rapidly evolving; no single standard yet |
| Data Poisoning | Corrupt training labels to degrade generalisation | Dataset curation and labelling pipeline | Robust loss functions, certified data cleansing, influence functions | Theoretical maturity; tooling maturing |
| Supply Chain Attack | Compromised pre-trained weights or libraries | Model zoos, hub repositories | Weight hashing, provenance signing, SBOM for AI | Early adoption phase (MLOps-integrated) |
| Side-Channel Leakage | Timing/power analysis reveals model architecture or data | Hardware inference accelerators (GPU/TPU) | Constant-time inference, hardware attestation, TEE deployment | Research stage; critical infrastructure focus |
| Trojan / Latent Backdoor | Hidden trigger activates misclassification at deployment | Fine-tuned or transfer-learned models | Neural Cleanse, STRIP, Activation Clustering detection | Tooling available; integration limited |
| Gradient Leakage | Reconstruction of training data from shared gradients (FL) | Federated learning aggregation server | Secure aggregation, gradient compression, DP in FL | Standardised in privacy-sensitive FL deployments |

## 9. A responsible AI deployment Lifecycle Framework.

All of the challenges discussed in this chapter drive towards the need to conceive a lifecycle approach to responsible AI deployment wherein ethical, security and accountability concerns are coupled with the phases of the system development and operation cycle as opposed to being discussed as an afterthought or compliance requirement that must be fulfilled at any given point in time. The lifecycle view presented in this paper consists of seven stages, which are problem framing and stakeholder engagement, data governance and preparation, model development and fairness-conscious training, security testing and robustness testing, transparency and explainability merger, deployment and observational tracking, and decommissioning and impact evaluation.

Responsible deployment is based on problem framing, which is often not given much effort at practical levels. This stage will entail identifying the social issue to be solved, surveying the stakeholder environment, defining the decision environment of which the model will be applied, and pinpointing possible harms and their incidence among the affected populations, and deciding whether a deep learning solution is suitable or whether some more understandable solutions would work better in the context of

deployment. It has been demonstrated that participatory design practices, community consultations and impact assessment done in combination with domain experts and communities impacted reveal ethical hazards and design expectations which the purely technical group will consistently miss.

A vital and frequently overlooked responsible deployment element is operational monitoring the chronic assessment of the behaviour of a deployed model in comparison with pre-established performance, fairness, and safety goals. Unlike classical software, deep learning models have a distributional drift, that is, as in production the statistical properties of the inputs received by the model change gradually, and its resultant performance deteriorates, which can be slow and hard to observe without an explicit monitoring instrumentation. Data drift detection, model performance monitoring, and alerting are now routinely implemented on MLOps platforms, under the assumption that such pipelines will now also incorporate fairness metrics and uncertainty calibration monitoring—although implementation of all of these underpinning principles is still early. The regulatory condition on post deployment monitoring and report of incidents as the EU AI Act on high-risk systems provide are establishing institutional incentives to encourage organisations to invest in such capabilities.

## 10. Future Research Preferences and Unsolved Problems.

There are a number of frontier research directions which will dramatically transform the picture of ethical, secure, and responsible implementation of deep learning systems in the near future. Human-feedback-based reinforcement learning (RLHF), direct preference optimisation (DPO), and constitutional AI are the three development directions that demonstrate a paradigm shift in the process of alignment of large language models and foundation models with human values (Bai et al., 2022). The methods aim at encoding the moral principles directly into the model reward or preference model and not merely use post-hoc filtering behaviours or external governance systems. Although there are positive empirical findings, some underlying questions exist concerning the stability of aligning behaviour during distributional shift, (1) how perspective in alignment techniques is susceptible to jail breaking as well as predatory (adversarial) fine-tuning and (2) whether human feedback signals during the orientation of alignment are culturally specific.

Mechanistic interpretability - the project of determining the internal computational processes the neural networks use to generate their behaviour, on the level of circuits, features and sub modules - is becoming an elementary research direction in both security and accountability. Experiments by the interpretability team and other researchers have shown that it is possible to identify a few specific circuits in transformer language models that are related to named entity recognition, indirect object identification, or

factual recall (Elhage et al., 2021; Olsson et al., 2022), which in turn allows considering a set of interventions: patching, ablating, and steering the model behaviourors to change them in principled ways. In case the idea of mechanistic interpretability progresses at a consistent pace, it can eventually be used to achieve formal verification of neural networks behaviour, bridging the gap between the certification mechanisms used on safety-critical software today and those in use today with deep learning.

A direction towards systems that are more capable and understandable than both paradigms alone has been proposed by neuro-symbolic integration, that integrates the limitations of deep learning with the logical reasoning and compositionality of symbolic AI. Neural Theorem Provers, neural logic programs, and neural logic programs with an added formal reasoning engine all are examples of architectures which combine neural and symbolic components and can be seen as complimentary to the subsymbolic feature learning of the neural one. How such hybrid systems can be aligned to formal regulatory demands of explainability and traceability can be found to be more manageable than purely neural based engines can be, which points to neuro-symbolic integration as a promising design strategy in highly stakes regulated use cases.

Convergence of responsible AI and environmental sustainability is gaining increased interest due to the rise of carbon footprint of training and deployment of large-scale deep learning systems being the matter of public and regulatory interest. Model compression methods have been shown to minimise the energy used to make an inference, pruning, quantisation, knowledge distillation and neural architecture search are methods that do not only minimise the power used during inference but its capacity to memorise training data can furthermore lead to co-benefits to privacy. The idea of green AI (Schwartz et al., 2020) suggests that the computational efficiency and accuracy measures should be reported in research papers, and how the lifecycle environmental impact of AI should be analyzed are pointing their way into the AI deployment governance processes. The conflict between the scale which enables the improvement of the capabilities of the foundation models and the ecological costs of the said scale will become a staple challenge of the responsible AI development over the next decade.

## 11. Conclusion

This chapter has introduced the treatment of the ethical, security, and responsibility facets in a holistic manner that should inform the practice of the deep learning systems in real-life applications. The main paper theme is that trustworthiness in AI is not a post-hoc property either imported into a system or patented on to a system but that it should be designed into each phase of the system lifecycle, including problem formulation and data controls as well as on system design, security testing, implementation, and monitoring of use. The technical issues are also significant and unsolved: certified

adversarial defences are scale-limited and of small perturbation radii; fairness definitions are mathematically incompatible in non-trivial models; differential privacy has an accuracy cost that limits its use in high-complexity systems; the mechanistic nature of formal verification of neural network behaviour is immature.

Nevertheless, there is no denying that the trend of research and regulatory development is clearly towards more principled, evidence-based, institutionally accountable AI implementation. The codification of the EU AI Act, NIST AI RMF, and a more and more diverse range of industry-specific technical standards give a regulatory space in which responsible deployment can be implemented an ever more tangible form. The coming of age of privacy-centric machine learning, adversarial resilience, and explainability as fields of engineering - with benchmarking, open-source infrastructure, and guidance - is mitigating the obstacles to the implementation of trustworthiness in the production systems. Mechanistic interpretability, the causal AI, and neuro-symbolic integration are convergent, and it holds the key to a better understanding of the neural network behaviour that can yield the formal assurance in the high-stakes applications.

The practice of responsible deployment of deep learning will ultimately demand more than the technical innovation; it will demand institutional commitment, interdisciplinary cooperation and actual interaction between the social and political environment in which an AI system must be used. Each of the researchers, practitioners, policymakers and the affected communities bring in invaluable insights to the challenge and composite integration of the insights is the hallmark work of the responsible AI enterprise. Chapter has been striving to offer to researchers and practitioners a spurring idea and conceptual as well as an empirical basis of approaching that task and to put the technical aspects of trustworthy AI in the wider context of ethical and government systems that give them their final meaning and urgency.58

# Question Bank

1) What is meant by trustworthy deep learning?
2) Why is trustworthiness considered a critical requirement in modern AI systems?
3) What are the three core pillars of trustworthy deep learning?
4) Explain the concept of robustness in deep learning.
5) What is uncertainty quantification in neural networks?
6) Define adversarial resilience in machine learning systems.
7) Why is benchmark accuracy insufficient for real-world AI deployment?
8) What is the deployment gap in deep learning systems?
9) Discuss real-world failures of deep learning in healthcare.
10) How do adversarial examples expose weaknesses in neural networks?
11) What is distributional shift and why is it problematic?
12) Explain the i.i.d. assumption in machine learning.
13) Why does the i.i.d. assumption fail in real-world scenarios?
14) What are the consequences of model miscalibration?
15) Define epistemic uncertainty.
16) Define aleatoric uncertainty.
17) How does uncertainty affect decision-making in high-stakes applications?
18) What is overconfidence in neural networks?
19) Why is interpretability important in AI systems?
20) Differentiate between interpretability and explainability.
21) What are the ethical concerns associated with deep learning?
22) Explain fairness in machine learning.
23) What is demographic bias in AI models?
24) How can training data lead to discrimination in AI systems?
25) What are spurious correlations in deep learning?
26) Explain the role of regulatory frameworks in AI governance.
27) What is the EU AI Act and its relevance to AI systems?
28) Why is accountability important in AI deployment?
29) What are high-risk AI systems?
30) Explain the concept of model governance.

31) What is a feedforward neural network?
32) Explain the concept of backpropagation.
33) What are activation functions in neural networks?
34) What is a loss function?
35) Define gradient descent optimization.
36) What is overfitting in deep learning?
37) What is underfitting?
38) Explain generalization in neural networks.
39) What is a convolutional neural network (CNN)?
40) What are residual networks?
41) Explain transformer architecture.
42) What is attention mechanism?
43) What are generative models?
44) What are the reliability issues in generative AI?
45) Define backdoor attacks in machine learning.
46) What is data poisoning?
47) Explain membership inference attacks.
48) What is machine unlearning?
49) What are privacy concerns in deep learning?
50) What is shortcut learning?
51) What is catastrophic forgetting?
52) What is aleatoric uncertainty?
53) What are the sources of aleatoric uncertainty?
54) What is heteroscedastic uncertainty?
55) Explain label noise and its impact on models.
56) What is epistemic uncertainty?
57) How is epistemic uncertainty estimated?
58) What are Bayesian neural networks?
59) What is posterior inference?
60) Explain deep ensembles.
61) What is Monte Carlo dropout?
62) How does MC dropout approximate Bayesian inference?
63) What is total predictive uncertainty?
64) Explain evidential deep learning.

65) What is conformal prediction?
66) How does conformal prediction ensure reliability?
67) What are uncertainty estimation challenges in large language models?
68) How is uncertainty used in deployment decisions?
69) What is Bayesian neural network theory?
70) Explain prior and posterior distributions in BNNs.
71) What is variational inference?
72) What is Markov Chain Monte Carlo (MCMC)?
73) Explain Laplace approximation.
74) What is stochastic weight averaging?
75) What are deep ensembles?
76) Why are ensembles effective for uncertainty estimation?
77) What is model diversity in ensembles?
78) Explain loss landscape geometry.
79) How do ensembles approximate Bayesian inference?
80) What are scalability challenges in probabilistic deep learning?
81) What are applications of probabilistic deep learning in healthcare?
82) How is uncertainty used in autonomous systems?
83) What is model calibration?
84) Define calibration curve.
85) What is miscalibration in neural networks?
86) Explain sharpness and resolution in calibration.
87) What are calibration metrics?
88) What is temperature scaling?
89) Explain post-hoc calibration methods.
90) What is confidence estimation?
91) How does distribution shift affect calibration?
92) What is group-wise calibration fairness?
93) Explain calibration in large language models.
94) What is robustness in deep learning?
95) Explain robustness-accuracy trade-off.
96) What are input corruptions?
97) What is stochastic noise?
98) What are natural perturbations?

99) Explain adversarial perturbations.
100) What is adversarial example phenomenon?
101) What are certified defenses?
102) What is randomized smoothing?
103) What is distribution shift?
104) Differentiate covariate shift and label shift.
105) What is concept drift?
106) What is domain generalization?
107) Explain invariant risk minimization.
108) What is distributionally robust optimization?
109) What is test-time adaptation?
110) What is adversarial machine learning?
111) Define white-box attacks.
112) Define black-box attacks.
113) What are gradient-based attacks?
114) What is PGD attack?
115) What are universal adversarial perturbations?
116) What are backdoor attacks?
117) Explain adversarial attacks in NLP systems.
118) What is adversarial training?
119) What are defense mechanisms against adversarial attacks?
120) Explain input preprocessing defenses.
121) What is adversarial detection?
122) What is transferability in adversarial attacks?
123) Why are adaptive attacks challenging?
124) What are evaluation metrics for trustworthy AI?
125) What is calibration error?
126) What is adversarial robustness metric?
127) What is fairness metric?
128) What is interpretability metric?
129) What are AI benchmarks?
130) What is RobustBench?
131) What is ImageNet-C?
132) What is benchmark saturation?

133) Explain train-test protocol standardization.
134) What are adversarial evaluation pipelines?
135) How is out-of-distribution performance measured?
136) What is ethical AI?
137) What are principles of responsible AI deployment?
138) What is fairness in AI systems?
139) What is accountability in AI?
140) What is transparency in AI?
141) What are privacy-preserving techniques?
142) What is differential privacy?
143) What is federated learning?
144) What are security threats in AI deployment?
145) What is adversarial resilience in deployment?
146) What is AI governance?
147) What are model cards?
148) What are datasheets for datasets?
149) What is algorithmic auditing?
150) What is human-AI trust?
151) What is automation bias?
152) What is algorithm aversion?
153) Explain applications of trustworthy AI in healthcare.
154) What are challenges of AI in autonomous systems?
155) How does AI impact financial systems?
156) What are risks of AI in natural language processing?
157) What are challenges in scientific AI applications?
158) What are risks in critical infrastructure AI systems?
159) Explain integrated approaches to trustworthy deep learning.
160) What are training-time robustness techniques?
161) What are post-hoc trustworthiness methods?
162) What is runtime monitoring in AI systems?
163) What are emerging trends in trustworthy AI?
164) What are foundation models?
165) What is transfer learning?
166) What is neuro-symbolic AI?

167) What are formal verification methods?

168) What is the privacy-utility trade-off?

169) What are open challenges in trustworthy deep learning?