Epistemic Artificial Intelligence is Essential for Machine Learning Models to Truly 'Know When They Do Not Know'

Shireen Kudukkil Manchingal[1*] **Andrew Bradley**[1] **Julian F. P. Kooij**[2]
Keivan Shariatmadar[3,4] **Neil Yorke-Smith**[5] **Fabio Cuzzolin**[1]

[1]School of Engineering, Computing and Mathematics, Oxford Brookes University, UK
[2]Cognitive Robotics, TU Delft, Netherlands [3]LMSD, Mechanical Engineering, KU Leuven
[4]Flanders Make@KU Leuven [5] STAR Lab, TU Delft, Netherlands
{smanchingal, abradley, fabio.cuzzolin}@brookes.ac.uk
{J.F.P.Kooij, n.yorke-smith}@tudelft.nl
{keivan.shariatmadar}@kuleuven.be

Abstract

Despite AI's impressive achievements, including recent advances in generative and large language models, there remains a significant gap in the ability of AI systems to handle uncertainty and generalize beyond their training data. AI models consistently fail to make robust enough predictions when facing unfamiliar or adversarial data. Traditional machine learning approaches struggle to address this issue, due to an overemphasis on data fitting, while current uncertainty quantification approaches suffer from serious limitations. This position paper posits a paradigm shift towards *epistemic artificial intelligence*, emphasizing the need for models to learn from what they know while at the same time acknowledging their ignorance, using the mathematics of *second-order uncertainty* measures. This approach, which leverages the expressive power of such measures to efficiently manage uncertainty, offers an effective way to improve the resilience and robustness of AI systems, allowing them to better handle unpredictable real-world environments.

1 Introduction

The success of artificial intelligence, especially deep learning [135], is indisputable. AI systems now perform many tasks at or above human levels, with generative models advancing into creativity [192], large language models (LLMs) [24] excelling in language manipulation, and significant advancements in multimodal AI [191]. However, this has also led to inflated expectations. For instance, autonomous vehicles have been touted as imminent breakthroughs for over a decade but technological challenges still remain a barrier to worldwide deployment [7, 248]. While generative models like ChatGPT create remarkable outputs, there is an increasing acknowledgment of the need to reassess AI's development path fundamentally [16]. Speculation about uncontrolled AI evolution and debates around artificial general intelligence (AGI) often overshadow pressing challenges AI must address today [202].

A significant limitation of current machine learning (ML) systems is their lack of robustness. Neural networks frequently make inaccurate and overconfident predictions when faced with uncertainties, such as out-of-distribution (OoD) samples, natural fluctuations, or adversarial disruptions [181, 103, 265]. These issues become safety-critical in autonomous vehicles due to models struggling to generalize across the diverse scenarios [86, 22] the vehicle may encounter. While efforts in overfitting mitigation [221, 165] and domain adaptation [89] are ongoing, these approaches are

*Corresponding author: *smanchingal@brookes.ac.uk*

Preprint. Under review.

arguably insufficient to address the fundamental challenges of robustness in a meaningful manner [94, 84, 23, 270].

There is a growing consensus that the accurate estimation of *uncertainty* [54] is vital to improve machine learning models' reliability [208, 120], with key applications to safety-critical areas such as autonomous driving [229], medical diagnosis [133], flood risk estimation [33], and structural health monitoring [242]. To fully capture the uncertainty in a system or process, it is necessary to recognize two main sources: *aleatoric* (predictable, irreducible) and *epistemic* (unpredictable, reducible) uncertainty. The former arises from randomness in the data; a simple example of this is the coin-toss, where the data generating process has a stochastic component that cannot be reduced by any additional source of information [109]. The latter, instead, arises from a *lack of knowledge* about the system. For example, the odds of drawing the Ace of spades at random from a deck of cards might be assumed to be 1/52. However, this is based upon a prior assumption that this is a complete, standard deck. An 'unknown' deck, however, may contain duplicates or missing cards, include jokers, or comprise multiple packs. Without this prior knowledge, the underlying model inevitably carries some uncertainty, which can be reduced with each subsequent observation. Hence, an awareness of the Socratic principle, to 'know that you do not know', is of paramount importance. The main source of uncertainty in AI (but also in its science and engineering applications) is indeed the lack of a sufficient amount of data to train a model, in both quantity and quality (*i.e.*, data fairly describing all regions of operation, including rare events). This uncertainty is epistemic in nature [55], as it concerns the model itself, and can be reduced by collecting more data or information.

While most scientists would agree that this is a profound problem, a defining issue for AI is *how* uncertainty should be managed, as existing uncertainty quantification (UQ) methods for AI have key limitations. *Bayesian* models are sensitive to prior mis-specification (with the risk of biasing the whole process) and incur heavy computational overhead [83, 29], while Bayesian Model Averaging (BMA) may dilute useful predictive information [105, 96]. *Ensemble* methods are computationally demanding [115, 102]. *Conformal* predictors primarily capture aleatoric uncertainty in a frequentist stance [15]. *Evidential* approaches violate asymptotic assumptions, struggle with out-of-distribution data [17, 238, 125, 222] and exhibit high inference times (§D).

This position paper advocates for a paradigm shift towards an Epistemic Artificial Intelligence emphasizing the importance of learning while acknowledging ignorance, using *second-order uncertainty measures* (§A) capable of overcoming those limitations thanks to their greater expressive power. **Epistemic AI rests on the 'paradoxical' principle that one should first and foremost learn from (or be ready for) the data it cannot see.** Prior to observing any data, the task at hand is thus completely unknown (albeit prior knowledge can be utilized to formalize the task and set a model space of solutions). The (limited) available evidence should only be used to temper our ignorance, to avoid 'catastrophically forgetting' how much we ignore about the problem.

Epistemic AI is supported by both theoretical arguments and strong empirical evidence (Sec. 4). **Firstly**, the use of second-order uncertainty measures allows Epistemic AI to explicitly represent model ignorance and properly account for uncertainty due to lack of knowledge without biasing the learning process, unlike traditional approaches (Sec. 4.1). **Secondly**, evidence is recently mounting that Epistemic AI can predict uncertainty more accurately, at lower inference times (§D), and more broadly outperform other UQ methods in terms of accuracy, robustness and calibration (Sec. 4.2). As a result, Epistemic AI is capable of reducing the likelihood of AI systems being 'surprised' by unexpected data or incapable to respond to unforeseen situations. This has enormous importance for mission-critical areas such as autonomous vehicles, or climate change and pandemic prediction, where long-term uncertainty is paramount as predictions concern the distant future and data is extremely scarce. Large language models learning 'epistemically' from data would be less likely to commit to false statements. Bias issues could be significantly mitigated, as epistemic models would not simply mimic the training data but account for possible future data. This paper presents arguments in support of Epistemic AI, discusses its potential and future challenges, while acknowledging alternative views.

Paper structure. Sec. 2 shows how estimating uncertainty aids robustness and adaptation. Sec. 3 reviews other models and perspectives. Sec. 4 introduces Epistemic AI, its theoretical (4.1) and empirical (4.2) support. Sec. 5.2 explores its potential future role in generative AI. Secs. 6–8 cover challenges, exciting opportunities in science and conclusions. Appendices §A and §C recall second-order measures and models. §B further details related work; §D provides additional results.

2 Why Uncertainty Quantification Matters

Adversarial Robustness. Traditional neural networks often suffer from overconfidence (softmax outputs reflect relative confidence, not true uncertainty) leading to high-confidence errors on out-of-distribution (OoD) or adversarially perturbed inputs [99, 103]. Fig. 1 compares a standard ResNet50 (*Traditional*) and an uncertainty-aware ResNet50 (*Epistemic*) [152] on ImageNet-A [104], an adversarially filtered dataset exposing model overconfidence. When both models, trained on ImageNet, are tested on ImageNet-A, the Traditional model remains highly confident in its misclassifications, whereas the Epistemic model assigns lower confidence to misclassifications, avoiding overconfidence.

Robustness and Domain Adaptation. Domain adaptation methods such as minimax learning [13], counterfactual error bounding [228], and custom loss functions use adversarial feature alignment for unsupervised adaptation [12, 6]. Reinforcement learning robustness employs adversarial strategies [185] and Bayesian Bellman formulations [72]. Out-of-distribution (OoD) and domain-generalization research leverages kernel methods [20, 73, 108, 161] and H-divergence–based adversarial learning [4], yet all falter under significant train–test shifts [198], where uncertainty management can improve adaptation and OoD detection [98, 124, 216].

Calibration. Neural networks are typically uncalibrated, *i.e.*, predicted confidence rarely equals accuracy [99]. Post-hoc techniques such as histogram and Bayesian binning, Platt scaling [173, 186], and regression extensions [128] improve this. Expected Calibration Error (ECE) [166], Adaptive CE [170], and loss-based adjustments [164, 144, 230] refine calibration but still overlook deeper uncertainty representation.

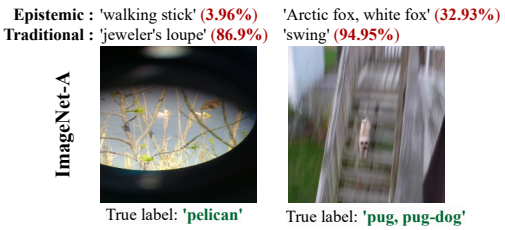


Figure 1: Confidence scores of uncertainty-aware (*Epistemic*) and (*Traditional*) model on ImageNet-A (adversarial). Unlike the epistemic model, traditional model is overconfident in misclassifications.

Sequential decision-making must also model the propagation of uncertainty, especially in safety-critical domains such as autonomous driving [231], where unmodeled perception or state uncertainty can cause compound errors and unsafe actions [207]. Effective quantification of epistemic uncertainty enables the system to detect unreliable predictions and act with 'human-like' cautiousness [120, 71].

3 Alternative Views

Traditional and Deterministic methods. Traditional models make deterministic predictions and lack uncertainty modeling, assuming exact input-output relations. Deep Deterministic Uncertainty (DDU) [163] estimates epistemic uncertainty via latent representation analysis or distance-sensitive functions rather than softmax probabilities [5, 257, 140, 163, 239]. However, regularization techniques like bi-Lipschitz, commonly used in these models, do not effectively improve OoD detection or calibration [189]. Unlike other methods, DDU captures uncertainty in the input space by detecting OoD samples rather than the prediction space. Both DDU and traditional models make point predictions (Fig. 2).

Bayesian Methods. Bayesian Deep Learning (BDL) [27, 145, 168] models network parameters as distributions using Bayesian neural networks (BNNs) [21, 88, 115], producing predictive distributions by sampling from an approximated posterior [109]. Despite advances in training via sampling [107, 169] and variational inference [21, 88, 121, 106, 87, 227, 201, 87], and successes in real-world tasks [242, 129], practical challenges remain. This includes the significant computational complexity associated with training [115] and inference (Tab. 1, §D), establishing appropriate prior distributions before training [255, 83], handling complex network architectures, and ensuring real-time applicability [163]. Furthermore, several studies have indicated that the use of single probability distributions to model epistemic lack of knowledge is, in fact, insufficient [29, 109].

Ensemble methods such as Deep Ensembles (DE) [132] and Epistemic Neural Networks (ENN) [176] estimate uncertainty by aggregating predictions from multiple models. DEs, in particular, have demonstrated strong performance in uncertainty estimation [177, 100, 1]. However, they are computationally intensive, with training and inference costs increasing linearly with the number of ensemble members, making them impractical for large models or real-time applications [140, 37, 102].

Conformal prediction [244] is a wrapper method applicable to any model, generating prediction sets (for accuracy guarantees) by computing empirical cumulative distributions and applying hypothesis testing to them. Several variants exist, *e.g.*, conditional, full conformal prediction [179, 178, 180,

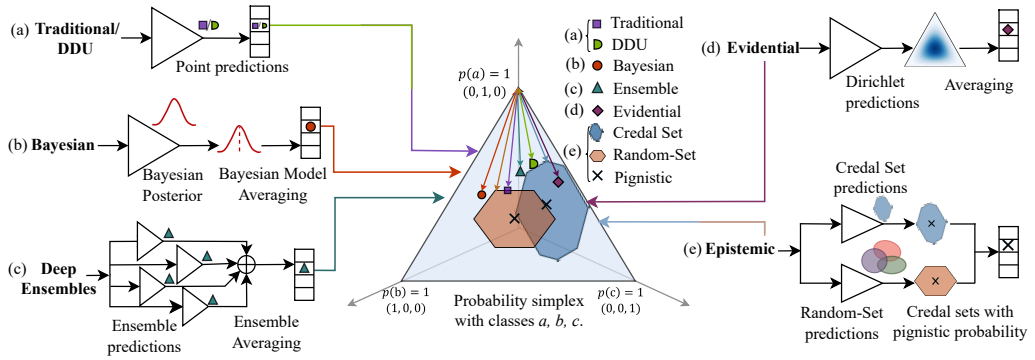


Figure 2: **Major approaches to uncertainty in AI.** Traditional networks and deterministic uncertainty models [163] (a) have fixed weights and output a deterministic value or probability vector. Bayesian neural networks [21, 88, 201] (b) estimate predictive distributions by integrating over posterior parameter distributions, often approximated via Bayesian Model Averaging (BMA) [105, 96] which averages predictions from sampled parameters. Similarly, deep ensembles [132] (c) average predictions from independently trained models. Evidential methods [209] (d) predict second-order Dirichlet parameters instead of softmax probabilities. Epistemic approaches (e) use second-order probability representations, such as interval probabilities, credal sets, or random-sets [152], with pignistic probabilities [218] derived from credal sets for comparison [150].

206, 171, 246, 243, 245, 28]. However, as it relies on building cumulative distribution functions of 'nonconformity scores' to which it applies classical hypothesis testing [15], conformal prediction basically models aleatoric, rather than epistemic uncertainty. Recent advances have been made towards an epistemic conformal learning, particularly under credal representations [139, 111].

Evidential Methods. The evidential framework [258] has been applied to neural networks [196], decision trees [78], K-nearest neighbours [66], and evidential deep learning classifiers for uncertainty quantification [233]. Sensoy et al. [209] introduced a Dirichlet-based classifier to estimate subjective opinions. While Dirichlet-based advances exist [147–149, 32], many loss functions fail to reduce epistemic uncertainty with more data, violating asymptotic assumptions [17]. Some methods rely on OoD training data, which may be unavailable or inadequate for robust detection [238], and even posterior networks with normalizing flows show limitations [125, 222].

Some critics, including ourselves, argue that classical probability theory cannot fully address 'second-level' uncertainty [109], suggesting the use of more generalized frameworks (§A), such as possibility theory [74], probability intervals [101], credal sets [137] or imprecise probabilities [247]. **The way epistemic uncertainty is managed and data is leveraged is, we feel, a defining issue for AI: with this position paper, we wish to contribute to this debate and indicate possible solutions through a paradigm shift which we term *epistemic artificial intelligence*.**

4 Epistemic Artificial Intelligence

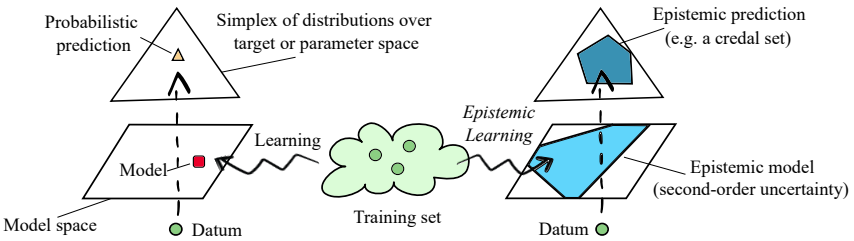


Figure 3: **Epistemic Learning.** Contrary to a traditional learning process in which a single model is learned from a training set to map new data to predictions, e.g. in the form of a probability distribution over the target space (left), epistemic learning outputs a second-order uncertainty measure (right).

As mentioned in Sec. 1, the core idea of Epistemic AI hinges on a paradox: the system must be designed not only to learn from the data it observes, but also to be prepared for data it has not yet encountered. The problem can be formalized as one of learning a mapping (*epistemic model*) from

input data points to predictions in the form of a second-order uncertainty measure (§A), either on the target space or on the parameter space of the model itself (Fig. 3). Later, this prediction may be updated in the light of new data. When the epistemic prediction is a credal set, as in most cases [150], a probability ('pignistic') estimate [218] can be computed as its center of mass (Fig. 2).

4.1 Why Epistemic AI is Essential

In addition to the limitations discussed in Sec. 3, existing models fundamentally struggle to capture epistemic uncertainty, primarily because a **single probability distribution cannot fully express ignorance** about the data-generating process [51]. Bayesian methods, in particular, though widely used, particularly falter in data-sparse or ambiguous settings because they must assign fixed belief mass even when knowledge is lacking. Uninformative priors such as Jeffreys' [112] are not invariant under reparameterization and can be improper, violating objectivity and the strong likelihood principle [214]. Moreover, priors must be specified even for systems without past data, leading to arbitrary modeling choices that can **bias the learning process for a long time** (Bernstein-von Mises theorem, [123]). Bayesian **posteriors may appear similar whether we have no knowledge or weak evidence**, conflating ignorance with imprecise belief and potentially causing misleading overconfidence. For example, as in the 'unknown' pack of cards scenario (Sec. 1), Bayesian inference treats uncertainty about the deck's composition as a single posterior distribution, rather than explicitly quantifying our lack of knowledge. Model selection and prior choice lack objective criteria and **prior sensitivity worsens with scarce data** [119, 18]. Further, Bayesian models **cannot naturally represent set-valued or propositional evidence**, because the additivity of probability forces allocation to individual outcomes, even when evidence supports sets of hypotheses, in opposition to random-sets which can naturally model missing data [51]. Bayes' rule also **assumes that new evidence is sharp and definitive**, which is unrealistic in many real-world cases. Hierarchical Bayesian models, which place priors over priors, can model epistemic uncertainty and potentially address some of these issues, but are very computationally expensive in high-dimensional or open-world settings.

Moreover, Bayesian inference tends to smooth out epistemic uncertainty by averaging over models, collapsing diverse possibilities into a single estimate and failing to distinguish knowns from unknowns [105, 96, 109]. Computationally, Bayesian models also often suffer from slow convergence and large inference times (Tab. 1, §D), limiting their suitability for real-time safety-critical systems like autonomous vehicles [115]. **We do not advocate for abandoning Bayesian approaches; rather, we argue that fully capturing epistemic uncertainty demands a generalization of Bayesian measures into broader, second-order frameworks (§A), calling for dedicated research and resource allocation toward these more expressive uncertainty models.**

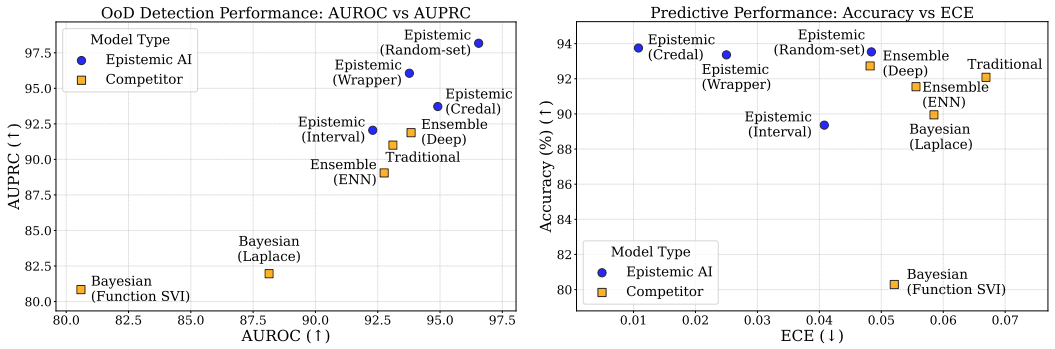
> Epistemic AI advocates for the adoption of second-order uncertainty measures, such as probability intervals, credal sets [136, 44] or random-sets, as they generalise classical probability using set-based representations and can richly encapsulate imprecision to model the epistemic uncertainty about an underlying, shifting data distribution, possibly the central challenge in machine learning.

Indeed, most second-order measures contain classical probability as a special case [54], with random-set reasoning subsuming Bayesian reasoning as a special case [213].

4.2 Empirical Support for Epistemic AI

Crucially, recent work on **Epistemic AI models** using second-order uncertainty measures (*Epistemic: Credal* [249], *Epistemic: Wrapper* [250], *Epistemic: Random-set* [152], *Epistemic: Interval* [252]) have demonstrated superior performance over **competitor models** (*Bayesian: Laplace* [106], *Bayesian: Function SVI* [201], *Ensemble: Deep* [132], *Ensemble: ENN* [176]) in classification tasks, based on experiments on large-scale benchmarks, including ImageNet, in terms of accuracy, robustness, uncertainty quantification and out-of-distribution detection [82], enhancing robustness in identifying novel or anomalous inputs.

In OoD detection, in particular, AUROC measures a model's ability to rank OoD samples above in-distribution ones by balancing true and false positive rates across thresholds, while AUPRC focuses on the scarce OoD class by summarizing precision-recall performance. High AUROC and AUPRC, respectively, indicate strong separability and reliable detection under class imbalance, making them complementary. In Fig. 4(a), Epistemic AI models (circles) cluster in the top-right on CIFAR-10 [194], demonstrating superior, consistent OoD detection due to their ability to preserve ignorance. In contrast, competitor methods (squares) perform worse and less consistently, empirically confirming



(a) (b)

Figure 4: Comparison of **Epistemic AI models** (circles) and **competitor models** (squares) on CIFAR-10. (a) **OoD detection performance (AUROC vs. AUPRC)**. Epistemic AI models cluster in the top-right (*high separability*) while competitor methods show a much greater spread (*lower performance*. (b) **Predictive performance (Accuracy vs. ECE).** Epistemic AI models cluster in the top-left (*high accuracy, low calibration error*) while competitor methods show poorer trade-offs (*weaker calibration*). Training details for all models are given in §D.

the theoretical advantage of Epistemic AI models based on second-order uncertainty measures under shifting data distributions. Moreover, Epistemic AI models also offer a better trade-off between accuracy and calibration. In Fig. 4(b), they dominate the top-left region, combining high accuracy with low Expected Calibration Error (ECE). More details on models, training/inference times (Tab. 1), and further evaluations (Fig. 8) can be found in §D.

Fig. 4(a) shows that second-order measures yield superior separation between in-distribution and out-of-distribution data. In Fig. 5, the Epistemic AI model (*Epistemic: Random-set*) is shown to exhibit low entropy for in-distribution and high entropy for out-of-distribution samples. This entropy gap (iD vs OoD entropy), reflected in both CIFAR-10 vs SVHN (left) and ImageNet vs ImageNet-O (right), demonstrates well-calibrated uncertainty estimates essential for reliable OoD detection and safer decision-making.

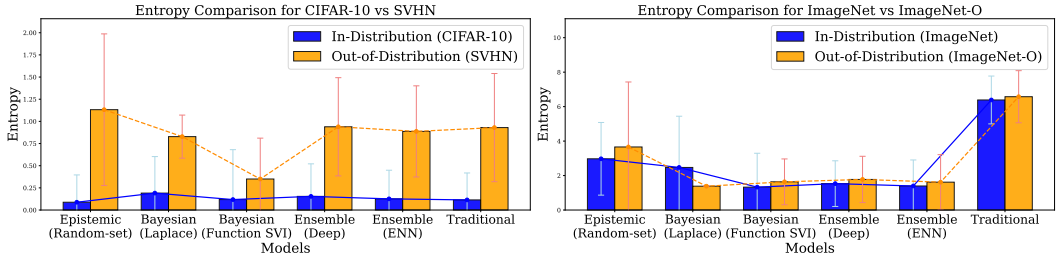


Figure 5: **Mean entropy** per model type for in-distribution (iD) and out-of-distribution (OoD) datasets, with error bars showing entropy standard deviation. The Epistemic AI model demonstrates better iD vs OoD entropy compared to other models.

5 Taking Epistemic AI Further

5.1 From Target-space to Parameter-space Representations

Epistemic uncertainty can be modelled at two levels: (i) a *target* level, where the network outputs an uncertainty measure on the target space, while its parameters (weights) remain deterministic; (ii) a *parameter* level, where uncertainty is modelled on the parameter space (*i.e.*, weights and biases).

The Epistemic AI models considered above are all **target-space** models. Credal-Set Interval Neural Networks (*Epistemic:Interval*) [252], based on Interval Neural Networks [172], predict probability intervals for classes. Credal Deep Ensembles (*Epistemic:Credal*) [249] use ensembles of credal networks to provide upper and lower probability bounds forming a credal set; trained with a distributionally-robust optimization (DRO)-inspired loss [131, 167, 203], CreDEs outperform DEs [132]. Credal Wrapper [250] (*Epistemic:Wrapper*) improves uncertainty estimation by 'wrapping' Bayesian and ensemble predictions as credal sets with upper/lower bounds per class, using the

'intersection probability' [43, 45, 53] to map a credal set to a single distribution. Random-Set Neural Networks (*Epistemic:Random-set*) [152] efficiently predict belief values for *sets* of classes, addressing ambiguity and incomplete data, using a budgeting method to reduce complexity.

A key future research direction is thus **extending Epistemic AI from target-level to parameter-level uncertainty representations, with the aim to fully generalize Bayesian (deep) learning**. One potential approach consists of transforming Bayesian Neural Network (BNN) posteriors into random-set posteriors without retraining. Using a transform proposed by Shafer [211], any likelihood or distribution can be converted into a random-set, efficiently represented as a Dirichlet distribution over parameter intervals [226]. Credal Bayesian Deep Learning [30], in opposition, introduces sets of posteriors over parameters, deriving predictive distributions at inference time that distinguish and quantify aleatoric and epistemic uncertainty, yielding either a set of outputs with guarantees or a single best prediction. Another promising direction is to employ Smets' Generalized Bayes Theorem (GBT) [217] (which produces belief functions [49], *i.e.*, finite random-sets, over parameters from a generalized likelihood and observations), under conditional cognitive independence (a generalization of i.i.d.), to directly learn random-set parametric representations from a training set.

Natural extensions to regression can also be envisaged. Credal Deep Ensembles (*Epistemic:Credal*) [249] may be applied to an ensemble of Bayesian regressors, each predicting a vertex of a credal prediction, with the final credal set as their convex hull. Random-Set Neural Networks (*Epistemic:Random-set*) [152] may also support regression, e.g., for object detection, by predicting Dirichlet distributions over Borel closed intervals [224] for bounding box coordinates. Class labels can be modeled as sets, enabling robust uncertainty in both spatial and categorical outputs.

5.2 An Epistemic Generative AI

An all-important effort is ongoing to extend the epistemic paradigm to generative AI.

Large Language Models (LLMs) [2, 8, 234] have shown strong performance in NLP tasks such as answering questions [118], reasoning [254], mathematical problem-solving [138], and code generation [200]. Pre-trained on large text corpora via next-token prediction, LLMs are fine-tuned for specific applications [38]. Despite their success, they face challenges like hallucinations [159]. Mechanisms to enhance their truthfulness (calibration) and quantify uncertainty could improve their reliability. Bayesian approaches such as Laplace-LORA [259], BLoB [253], and Monte-Carlo Dropout (MCD) [88], along with techniques like Bayes by Backprop (BBB) [21] and LORA Ensembles [14], have been applied to LLMs for uncertainty quantification. ENN-LLM [175] uses Epinet-inspired ensembles, while others leverage hidden states [34], softmax entropy [187] or semantic entropy [79]. These methods, however, often trade performance for inference efficiency.

An important challenge, both in the context of LLMs and beyond, is how to elicit second-order representations from 'traditional' ground truth datasets, such as question-answer pairs. ***How do we teach a model that the examples it sees are only samples from an incredibly rich set of possibilities?*** Developing appropriate evaluation methods for uncertainty-aware LLMs is another challenge that needs to be addressed before such models can be effectively trained and deployed. In the context of GenAI, **Epistemic AI can teach generative models the range of possible outputs they could produce from a limited training set**, capturing the epistemic uncertainty of the generative process itself. Our hypothesis is that modeling second-order uncertainty should enable generative models to better represent the *diversity of outputs*, particularly when training data is scarce or unrepresentative. For instance, allowing LLMs to predict probabilities for *sets of tokens* in a random-set framework, rather than single tokens, may allow them to capture a broader range of plausible outputs and improve overall accuracy. This could be especially useful for languages like Japanese or Arabic, where synonyms are prevalent and capturing a range of possible outputs is key to accurate predictions. Indeed, the random-set approach [152] to classification can be directly applied to Random-Set LLMs (RS-LLMs) [162], where belief functions over the vocabulary are predicted at each step instead of probability distributions, allowing language models to express ignorance. Hierarchical embedding can be used to cluster similar tokens into semantically-meaningful focal sets; sentence uncertainty can then be calculated as the mean credal width of its tokens. Random-set methods can also extend to generative AI via inferential models [155, 156], rooted in Dempster's belief-function theory [61] and Fisher's fiducial inference [183]. These models can infer belief functions over neural network weights, treating generative models like GANs as auxiliary equations. Gaussian noise can be transformed into a predictive random-set, generating output variability beyond traditional methods.

6 Challenges

Epistemic AI is effective and a potential key to addressing fundamental issues in machine learning. Still, challenges that are shared across uncertainty quantification may be exacerbated when using second-order uncertainty measures, owing to their higher expressiveness and complexity.

Applying second-order uncertainty measures to machine learning. Working with sets of distributions (*e.g.*, credal or random-sets) may involve costly sampling and inference procedures, particularly for decision-making [11, 10]. Recent work has addressed this by employing set budgeting techniques to efficiently constrain the complexity of using random-sets [152]. However, further research is needed to expand this to other second-order representations. Evaluation is an outstanding problem, as standard metrics do not apply directly to epistemic predictions. To address this, a unified framework to compare predictions across Bayesian, credal, random-set, ensemble, and evidential models was recently introduced in [150]. Nevertheless, an accepted global metric for comparing uncertainty-aware predictions is still wanting.

Scaling up. Most evidential approaches [209] struggle with scalability beyond medium-sized datasets. The clustering approach in the random-set approach [152] has unlocked the potential of random-set representations to large datasets like ImageNet and architectures like Vision Transformers, with future extensions possibly incorporating Dirichlet mixture models [260] and dynamic clustering [210] for continual learning. A key challenge remains: *can epistemic representations scale to foundation models and massive datasets?* While efficient belief function/random-set representations have been explored [195], further work is needed. Quantum approaches show some promise, with recent work on belief representation [269], combination [268], and integration into quantum circuits [256].

From one-off to continual learning. *Continual learning* is a more faithful representation of life-long real-world learning processes, especially in contexts in which models are continually updated in the light of streaming data whose distribution, however, may vary over time in unknown ways. Most research has focused on supervised learning and preventing models from 'forgetting' [122], using priors, task-specific parameters, or replay buffers [197]. Recently, unsupervised and semi-supervised settings, such as domain-incremental learning [240], have gained attention. Online learning and convex optimization [56] offer robustness guarantees by minimizing regret. Despite recent efforts [266, 113], a unified framework linking uncertainty modeling and continual learning remains an entirely open challenge, not just for Epistemic AI but for uncertainty quantification in general.

Learning and symbolic reasoning under uncertainty. Epistemic uncertainty can be reduced by collecting more data or incorporating prior knowledge, such as symbolic information (*e.g.*, Snorkel [193]), but data alone does not guarantee better performance, as seen in autonomous vehicle failures. *Neurosymbolic AI* integrates symbolic reasoning with deep learning to regularize predictions and enable knowledge transfer across domains [77, 154]. Current NeurAI frameworks enforce symbolic constraints but struggle with assessing output frequency or scaling to large knowledge bases [3]. Approaches like DeepProbLog [153] and DL2 [81] leverage fuzzy and probabilistic semantics but lack epistemic uncertainty modeling. Potential solutions include designing epistemic semantic losses or using logical circuits like trigger graphs [237] to extend DeepProbLog-style reasoning.

Statistical guarantees. Most current Epistemic AI methods do not provide statistical coverage guarantees on their predictions, albeit they can do so in combination with classical conformal learning [151]. Already mentioned efforts to generalise conformal learning certainly go in this direction. Recent studies have been looking at extending the notion of confidence interval to belief functions, under the name of confidence structures [69], which generalise standard confidence distributions and generate 'frequency-calibrated' belief functions. Also in the random-set setting, Inferential Models (IMs) can produce belief functions with well-defined frequentist properties [156]. An alternative approach relies on the notion of 'predictive' belief function [65], which, under repeated sampling, is less committed than the true probability distribution of interest with some prescribed probability.

7 Opportunities: Epistemic AI for Science and Engineering

Alongside challenges, Epistemic AI also presents a golden opportunity to enhance the AI-driven revolution in fields such as drug discovery, materials science, and astronomy. For example, DeepMind's Alphafold [116] revolutionized protein structure prediction, impacting molecular biology. Still, models like Alphafold and those used in weather forecasting often fail to model uncertainty in their predictions, which is crucial in real-world applications such as climate change, additive manufacturing or modeling of **nuclear fusion** plasma. The recently open-sourced Alphafoldv3 and

neural operator (NO) models [146, 90, 271], such as those used in nuclear fusion and climate prediction, show promise but need better uncertainty quantification to improve accuracy and efficiency, particularly in complex systems like differential equations.

The potential of **neural operators** in Epistemic AI, as powerful surrogate models for solving PDE-governed systems across science and engineering, is significant. However, despite recent work on Bayesian [146] and conformal prediction (CP) [97], NOs struggle with uncertainty due to limited data or PDE misspecification. CP provides calibrated uncertainty but needs additional calibration data, which can be costly. As neural operators learn from data a functional mapping between input and output functions (e.g., the boundary conditions and the solutions of a system of differential equations), applying epistemic learning to them involves solving the problem of **quantifying uncertainty in functional spaces**, generalising the classical neural network treatment. Epistemic AI can help model uncertainty in the gap between low- and high-fidelity simulations, as shown in fusion plasma edge modeling [80]. It can also enable the robust treatment of parameter uncertainty: for instance, finite element method (FEM) simulations often estimate physical parameters within confidence bounds. Given the breadth of NO applications, from climate to materials science, the impact of epistemic methods is potentially profound.

A paramount use case scenario is **climate change**, which is altering the weather cycle at global scale, amplifying extreme events like floods and droughts at continental scale [205]. Trends in the likelihood of extreme events, such as floods or droughts, are of particular interest to our society. An accurate prediction of climate change requires a correct representation of different compartments of the Earth system (*e.g.* atmosphere, ocean, and land) and the interactions between them. Each of these compartments is evolving and the interaction between them is highly dynamic. Some limited work exists on the possible use of AI for climate change, including prediction [223], mitigation [117] and adaptation [36, 35]. Interesting position papers and surveys on this have been published in recent years [41, 110]. Further, reliable long-term predictions require more than simple adaptation to a time series of data made available over time, highlighting the importance of quantifying epistemic uncertainty in the prediction of machine learning models trained on insufficient, sparse data to avoid forecasting errors and improve decision-making, with significant societal and scientific impact.

8 Conclusions

This position paper highlights the fact that **existing methods for uncertainty quantification in AI fail to efficiently model second-order uncertainty**, which is critical for epistemic uncertainty quantification and to give models the ability to 'truly' know when they do not know. We argued that there is a need for much further research in this area, not only in core machine learning but also in the context of generative AI and AI for science, highlighting the need for further research and testing to further develop this promising approach. We also pointed out that significant evidence is indeed starting to support the advantage of second-order uncertainty methods in machine learning.

Our **position is two-fold**: (a) We argue for the need to establish a concept we call **Epistemic AI**, according to which second-order uncertainty measures [52] (§A) are used to model epistemic uncertainty. The key argument is that ignorance is better represented through second-order uncertainty measures, which capture the inherent uncertainty about unknowns. (b) While the computational challenges of Bayesian and Ensemble models have been widely recognized, the AI community has yet to fully explore alternative models that can efficiently estimate second-order uncertainty. We also note that while there has been some progress in areas like classification and regression, significant gaps remain in more complex tasks like GenAI and AI4Science. Moreover, critical questions remain about selecting the most appropriate model for second-order uncertainty estimation and understanding the broader challenges in scaling these methods for practical applications.

The recent breakthroughs in this area were made possible by realizing that it is not necessary to exploit the full expressive power of second-order uncertainty measures (§A) to achieve significant improvements. Effective scalability can be attained by designing structures rich enough to harness the representational potential of second-order uncertainty measures while remaining computationally feasible. This can be achieved, e.g., through a suitable collection of focal sets [152], lower/upper probability structures [251], or a fixed budget of vertices for credal representations. Building on these results, a more principled and systematic exploration of these structures is now necessary to fully realize the vision of this paper.

An exciting future research direction would be the formal definition and study of specific families of random-sets, analogous to the families of probability distributions in classical probability (Gamma,

exponential, etc.), leading to more efficient and scalable computational models and driving further AI advances. The integration of Epistemic AI with continual, neurosymbolic and neural operator learning poses a set of exciting challenges moving forward.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. 964505 (E-pi). We thank the entire team of E-pi for the insightful discussions: Matthijs Spaan, Hans Hallez, David Moens, Maryam Sultana, Guopeng Li, Moritz Zanger, Pascal van der Vaart, Noah Schutte, Muhammad Mubashar, Kaizheng Wang, Adam Faza.

References

[1] Abe, T., Buchanan, E. K., Pleiss, G., Zemel, R., and Cunningham, J. P. (2022). Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660.

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[3] Aditya, S., Yang, Y., and Baral, C. (2019). Integrating knowledge and reasoning in image understanding. *arXiv preprint arXiv:1906.09954*.

[4] Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. (2019). Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*.

[5] Alemi, A. A., Fischer, I., and Dillon, J. V. (2018). Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*.

[6] Alijani, S., Fayyad, J., and Najjaran, H. (2024). Vision transformers in domain adaptation and domain generalization: a study of robustness. *Neural Computing and Applications*, 36(29):17979–18007.

[7] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[8] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

[9] Antonucci, A. and Cuzzolin, F. (2010). Credal sets approximation by lower probabilities: application to credal networks. In *Computational Intelligence for Knowledge-Based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28-July 2, 2010. Proceedings 13*, pages 716–725. Springer.

[10] Augustin, T. (2022). Statistics with imprecise probabilities—a short survey. *Uncertainty in Engineering Introduction to Methods and Applications*, 67.

[11] Augustin, T., Coolen, F. P., De Cooman, G., and Troffaes, M. C. (2014). *Introduction to imprecise probabilities*, volume 591. John Wiley & Sons.

[12] Awais, M., Zhou, F., Xu, H., Hong, L., Luo, P., Bae, S.-H., and Li, Z. (2021). Adversarial robustness for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8568–8577.

[13] Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR.

[14] Balabanov, O. and Linander, H. (2024). Uncertainty quantification in fine-tuned llms using lora ensembles. *arXiv preprint arXiv:2402.12264*.

[15] Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

[16] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

[17] Bengs, V., Hüllermeier, E., and Waegeman, W. (2022). Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35:29205–29216.

[18] Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

[19] Bezdek, J. C., Keller, J., Krisnapuram, R., and Pal, N. (1999). *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4. Springer Science & Business Media.

[20] Blanchard, G., Lee, G., and Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

[21] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

[22] Bojarski, M. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

[23] Braiek, H. B. and Khomh, F. (2025). Machine learning robustness: A primer. In *Trustworthy AI in Medical Imaging*, pages 37–71. Elsevier.

[24] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[25] Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.

[26] Bueno, T. P., Mauá, D. D., Barros, L. N., and Cozman, F. G. (2017). Modeling markov decision processes with imprecise probabilities using probabilistic logic programming. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pages 49–60. PMLR.

[27] Buntine, W. L. and Weigend, A. S. (1991). Bayesian back-propagation. *Complex Syst.*, 5.

[28] Campos, M., Farinhas, A., Zerva, C., Figueiredo, M. A., and Martins, A. F. (2024). Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

[29] Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., and Lee, I. (2023). Imprecise Bayesian neural networks. *arXiv preprint arXiv:2302.09656*.

[30] Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., and Lee, I. (2024a). Credal bayesian deep learning. *Transactions on Machine Learning Research*.

[31] Caprio, M., Sultana, M., Elia, E., and Cuzzolin, F. (2024b). Credal learning theory. *arXiv preprint arXiv:2402.00957*.

[32] Charpentier, B., Zügner, D., and Günnemann, S. (2020). Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367.

[33] Chaudhary, P., Leitão, J. P., Donauer, T., D'Aronco, S., Perraudin, N., Obozinski, G., Perez-Cruz, F., Schindler, K., Wegner, J. D., and Russo, S. (2022). Flood uncertainty estimation using deep ensembles. *Water*, 14(19):2980.

[34] Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. (2024). Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

[35] Chen, L., Chen, Z., Zhang, Y., Liu, Y., Osman, A. I., Farghali, M., Hua, J., Al-Fatesh, A. S., Ihara, I., Rooney, D. W., and Yap, P. (2023). Artificial intelligence-based solutions for climate change: a review. *Environmental Chemistry Letters*.

[36] Cheong, S.-M., Sankaran, K., and Bastani, H. (2022). Artificial intelligence for climate change adaptation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12.

[37] Ciosek, K., Fortuin, V., Tomioka, R., Hofmann, K., and Turner, R. (2019). Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*.

[38] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[39] Corani, G., Antonucci, A., and Zaffalon, M. (2012). Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, pages 49–93.

[40] Corani, G. and Zaffalon, M. (2008). Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(4).

[41] Cowls, J., Tsamados, A., Taddeo, M., and Floridi, L. (2021). The ai gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *Ai & Society*, 38:283 – 307.

[42] Cozman, F. G. (2000). Credal networks. *Artificial Intelligence*, 120(2):199–233.

[43] Cuzzolin, F. (2007). Two new Bayesian approximations of belief functions based on convex geometry. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 37(4):993–1008.

[44] Cuzzolin, F. (2008). On the credal structure of consistent probabilities. In *European Workshop on Logics in Artificial Intelligence*, pages 126–139. Springer.

[45] Cuzzolin, F. (2009). The intersection probability and its properties. In Sossai, C. and Chemello, G., editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 5590 of *Lecture Notes in Computer Science*, pages 287–298. Springer, Berlin Heidelberg.

[46] Cuzzolin, F. (2010a). Credal semantics of Bayesian transformations in terms of probability intervals. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(2):421–432.

[47] Cuzzolin, F. (2010b). Three alternative combinatorial formulations of the theory of evidence. *Intelligent Data Analysis*, 14(4):439–464.

[48] Cuzzolin, F. (2011). On consistent approximations of belief functions in the mass space. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 287–298. Springer.

[49] Cuzzolin, F. (2014). *Belief functions: theory and applications*. Springer.

[50] Cuzzolin, F. (2018). Generalised max entropy classifiers. In *Belief Functions: Theory and Applications: 5th International Conference, BELIEF 2018, Compiègne, France, September 17-21, 2018, Proceedings 5*, pages 39–47. Springer.

[51] Cuzzolin, F. (2020). *The Geometry of Uncertainty: The Geometry of Imprecise Probabilities*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing.

[52] Cuzzolin, F. (2021). Uncertainty measures: The big picture. *arXiv preprint arXiv:2104.06839*.

[53] Cuzzolin, F. (2022). The intersection probability: betting with probability intervals. *arXiv preprint arXiv:2201.01729*.

[54] Cuzzolin, F. (2024). Uncertainty measures: A critical survey. *Information Fusion*, page 102609.

[55] Cuzzolin, F. and Sultana, M. (2024). *Epistemic Uncertainty in Artificial Intelligence*. Springer.

[56] Dall'Anese, E., Simonetto, A., Becker, S., and Madden, L. (2020). Optimization and learning with information streams: Time-varying algorithms and applications. *IEEE Signal Processing Magazine*, 37(3):71–83.

[57] De Campos, C. P. and Cozman, F. G. (2004). Inference in credal networks using multilinear programming. In *Proceedings of the Second Starting AI Researcher Symposium*, pages 50–61.

[58] de Finetti, B. (1974). *Theory of Probability*. Wiley, London.

[59] Delgado, K. V., De Barros, L. N., Cozman, F. G., and Shirota, R. (2009). Representing and solving factored markov decision processes with imprecise probabilities. *Proceedings ISIPTA, Durham, United Kingdom*, 18:61.

[60] Delgado, K. V., Sanner, S., and De Barros, L. N. (2011). Efficient solutions to factored mdps with imprecise transition probabilities. *Artificial Intelligence*, 175(9-10):1498–1527.

[61] Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30(2):205–247.

[62] Dempster, A. P. (2008). Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer.

[63] Denneberg, D. and Grabisch, M. (1999). Interaction transform of set functions over a finite set. *Information Sciences*, 121(1-2):149–170.

[64] Denœux, T. (1995). A k-nearest neighbor classification rule based on Dempster–Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813.

[65] Denœux, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252.

[66] Denœux, T. (2008). A k-nearest neighbor classification rule based on Dempster–Shafer theory. In Yager, R. R. and Liu, L., editors, *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219 of *Studies in Fuzziness and Soft Computing*, pages 737–760. Springer.

[67] Denoeux, T. (2021). Nn-evclus: Neural network-based evidential clustering. *Information Sciences*, 572:297–330.

[68] Denœux, T., Kanjanatarakul, O., and Sriboonchitta, S. (2015). Ek-nnclus: a clustering procedure based on the evidential k-nearest neighbor rule. *Knowledge-Based Systems*, 88:57–69.

[69] Denœux, T. and Li, S. (2018). Frequency-calibrated belief functions: review and new insights. *International Journal of Approximate Reasoning*, 92:232–254.

[70] Denœux, T. and Masson, M.-H. (2004). Evclus: evidential clustering of proximity data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):95–109.

[71] Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR.

[72] Derman, E., Mankowitz, D., Mann, T., and Mannor, S. (2020). A bayesian approach to robust reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 648–658. PMLR.

[73] Deshmukh, A. A., Lei, Y., Sharma, S., Dogan, U., Cutler, J. W., and Scott, C. (2019). A generalization error bound for multi-class domain generalization. *arXiv preprint arXiv:1905.10392*.

[74] Dubois, D. and Prade, H. (1990). Consonant approximations of belief functions. *International Journal of Approximate Reasoning*, 4:419–449.

[75] Dubois, D. and Prade, H. (2012). *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science & Business Media.

[76] Dupuis, B., Viallard, P., Deligiannidis, G., and Simsekli, U. (2024). Uniform generalization bounds on data-dependent hypothesis sets via pac-bayesian theory on random sets. *arXiv preprint arXiv:2404.17442*.

[77] d'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-symbolic learning systems*. Springer.

[78] Elouedi, Z., Mellouli, K., and Smets, P. (Madrid, 2000). Decision trees using the belief function theory. In *Proceedings of the Eighth International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2000)*, volume 1, pages 141–148.

[79] Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

[80] Faza, G. A., Shariatmadar, K., Hallez, H., and Moens, D. (2024). Interval reduced order surrogate modelling framework for uncertainty quantification. In *AIAA Scitech 2024 Forum*, page 0387.

[81] Fischer, M., Balunovic, M., Drachsler-Cohen, D., Gehr, T., Zhang, C., and Vechev, M. (2019). Dl2: training and querying neural networks with logic. In *International Conference on Machine Learning*, pages 1931–1941. PMLR.

[82] Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081.

[83] Fortuin, V. (2022). Priors in Bayesian Deep Learning: A Review. *International Statistical Review*, 90.

[84] Freiesleben, T. and Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109.

[85] Fukuda, K. and Prodon, A. (1995). Double description method revisited. In *Franco-Japanese and Franco-Chinese conference on combinatorics and computer science*, pages 91–111. Springer.

[86] Fursa, I., Fandi, E., Musat, V., Culley, J., Gil, E., Teeti, I., Bilous, L., Sluis, I. V., Rast, A., and Bradley, A. (2022). Worsening perception: Real-time degradation of autonomous vehicle perception performance for simulation of adverse weather conditions. *SAE Intl. J CAV 5(1):87-100*.

[87] Gal, Y. and Ghahramani, Z. (2015). Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.

[88] Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

[89] Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

[90] Garg, S. and Chakraborty, S. (2023). Vb-deeponet: A bayesian operator learning framework for uncertainty quantification. *Engineering Applications of Artificial Intelligence*, 118:105685.

[91] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.

[92] Goan, E. and Fookes, C. (2020). Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87.

[93] Gong, W. and Cuzzolin, F. (2017). A belief-theoretical approach to example-based pose estimation. *IEEE Transactions on Fuzzy Systems*, 26(2):598–611.

[94] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[95] Grabisch, M., Sugeno, M., and Murofushi, T. (2000). *Fuzzy measures and integrals: theory and applications*. New York: Springer.

[96] Graefe, A., Küchenhoff, H., Stierle, V., and Riedl, B. (2015). Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951.

[97] Gray, A., Gopakumar, V., Rousseau, S., and Destercke, S. (2025). Guaranteed confidence-band enclosures for pde surrogates. *arXiv preprint arXiv:2501.18426*.

[98] Gulrajani, I. and Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

[99] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

[100] Gustafsson, F. K., Danelljan, M., and Schon, T. B. (2020). Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319.

[101] Halpern, J. Y. (2017). *Reasoning About Uncertainty*. MIT Press.

[102] He, B., Lakshminarayanan, B., and Teh, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022.

[103] Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

[104] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

[105] Hinne, M., Gronau, Q. F., van den Bergh, D., and Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215.

[106] Hobbhahn, M., Kristiadi, A., and Hennig, P. (2022). Fast predictive uncertainty for classification with Bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pages 822–832. PMLR.

[107] Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

[108] Hu, S., Zhang, K., Chen, Z., and Chan, L. (2020). Domain generalization via multidomain discriminant analysis. In *Uncertainty in artificial intelligence*, pages 292–302. PMLR.

[109] Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

[110] Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14.

[111] Javanmardi, A., Stutz, D., and Hüllermeier, E. (2024). Conformalized credal set predictors. *Advances in Neural Information Processing Systems*, 37:116987–117014.

[112] Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

[113] Jha, S., Gong, D., Zhao, H., and Yao, L. (2024). Npcl: Neural processes for uncertainty-aware continual learning. *Advances in Neural Information Processing Systems*, 36.

[114] Jiao, L., Denœux, T., Liu, Z.-G., and Pan, Q. (2022). Egmm: An evidential version of the gaussian mixture model for clustering. *Applied Soft Computing*, 129:109619.

[115] Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48.

[116] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

[117] Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., and Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12:518 – 527.

[118] Kalla, D., Smith, N., Samaah, F., and Kuraku, S. (2023). Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

[119] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association*, 91(435):1343–1370.

[120] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *arXiv:1703.04977*.

[121] Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local Reparameterization trick. *Advances in Neural Information Processing Systems*, 28.

[122] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[123] Kleijn, B. J. and Van der Vaart, A. W. (2012). The bernstein-von-mises theorem under misspecification.

[124] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

[125] Kopetzki, A.-K., Charpentier, B., Zügner, D., Giri, S., and Günnemann, S. (2021). Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR.

[126] Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge university press.

[127] Krishnapuram, R. and Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, 1(2):98–110.

[128] Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

[129] Kwon, Y., Won, J.-H., Kim, B. J., and Paik, M. C. (2020). Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816.

[130] Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. (2021). Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

[131] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.

[132] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30.

[133] Lambrou, A., Papadopoulos, H., and Gammerman, A. (2010). Reliable confidence measures for medical diagnosis with evolutionary algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 15(1):93–99.

[134] Lampinen, J. and Vehtari, A. (2001). Bayesian approach for neural networks—review and case studies. *Neural networks*, 14(3):257–274.

[135] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[136] Levi, I. (1980a). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press.

[137] Levi, I. (1980b). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. The MIT Press, Cambridge, Massachusetts.

[138] Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. (2022). Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

[139] Lienen, J., Demir, C., and Hüllermeier, E. (2023). Conformal credal self-supervised learning. In *Conformal and Probabilistic Prediction with Applications*, pages 214–233. PMLR.

[140] Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. (2020). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512.

[141] Liu, Z.-G., Dezert, J., Mercier, G., and Pan, Q. (2012). Belief c-means: An extension of fuzzy c-means algorithm in belief functions framework. *Pattern Recognition Letters*, 33(3):291–300.

[142] Liu, Z.-G., Liu, Y., Dezert, J., and Cuzzolin, F. (2019). Evidence combination based on credal belief redistribution for pattern classification. *IEEE Transactions on Fuzzy Systems*, 28(4):618–631.

[143] Liu, Z.-g., Pan, Q., Dezert, J., and Mercier, G. (2015). Credal c-means clustering method based on belief functions. *Knowledge-based systems*, 74:119–132.

[144] Luo, R., Bhatnagar, A., Bai, Y., Zhao, S., Wang, H., Xiong, C., Savarese, S., Ermon, S., Schmerling, E., and Pavone, M. (2022). Local calibration: metrics and recalibration. In *Uncertainty in Artificial Intelligence*, pages 1286–1295. PMLR.

[145] MacKay, D. J. C. (1992). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472.

[146] Magnani, E., Krämer, N., Eschenhagen, R., Rosasco, L., and Hennig, P. (2022). Approximate bayesian neural operators: Uncertainty quantification for parametric pdes. *arXiv preprint arXiv:2208.01565*.

[147] Malinin, A. and Gales, M. (2018). Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

[148] Malinin, A. and Gales, M. (2019). Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

[149] Malinin, A., Mlodozeniec, B., and Gales, M. (2019). Ensemble distribution distillation. In *International Conference on Learning Representations*.

[150] Manchingal, S. K., Mubashar, M., Wang, K., and Cuzzolin, F. (2025a). A unified evaluation framework for epistemic predictions.

[151] Manchingal, S. K., Mubashar, M., Wang, K., Shariatmadar, K., and Cuzzolin, F. (2024). Random-set neural networks (rs-nn).

[152] Manchingal, S. K., Mubashar, M., Wang, K., Shariatmadar, K., and Cuzzolin, F. (2025b). Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*.

[153] Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 31.

[154] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., and Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

[155] Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313.

[156] Martin, R. and Liu, C. (2015). *Inferential models: reasoning with uncertainty*. CRC Press.

[157] Masson, M.-H. and Denoeux, T. (2008). Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397.

[158] Masson, M.-H. and Denoeux, T. (2009). Recm: relational evidential c-means algorithm. *Pattern Recognition Letters*, 30(11):1015–1026.

[159] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

[160] Molchanov, I. (2017). Random sets and random functions. *Theory of Random Sets*, pages 451–552.

[161] Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–10–I–18. JMLR.org.

[162] Mubashar, M., Manchingal, S. K., and Cuzzolin, F. (2025). Random-set large language models. *arXiv preprint arXiv:2504.18085*.

[163] Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. (2023). Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.

[164] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. (2020). Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299.

[165] Muşat, V., Fursa, I., Newman, P., Cuzzolin, F., and Bradley, A. (2021). Multi-weather city: Adverse weather stacking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2915.

[166] Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

[167] Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

[168] Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

[169] Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

[170] Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

[171] Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385–392. Morgan Kaufmann.

[172] Oala, L., Heiß, C., Macdonald, J., März, M., Kutyniok, G., and Samek, W. (2021). Detecting failure modes in image reconstructions with interval neural network uncertainty. *International Journal of Computer Assisted Radiology and Surgery*, 16:2089–2097.

[173] Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., and Ziegler, A. (2023). Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, 42(29):5451–5478.

[174] Oren, Y., Spaan, M. T., and Böhmer, W. (2022). E-mcts: Deep exploration in model-based reinforcement learning by planning with epistemic uncertainty. *arXiv preprint arXiv:2210.13455*.

[175] Osband, I., Asghari, S. M., Van Roy, B., McAleese, N., Aslanides, J., and Irving, G. (2022). Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568*.

[176] Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahimi, M., Lu, X., and Van Roy, B. (2024). Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36.

[177] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

[178] Papadopoulos, H., Gammerman, A., and Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, AIA '08, page 64–69, USA. ACTA Press.

[179] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002a). Inductive confidence machines for regression. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg. Springer Berlin Heidelberg.

[180] Papadopoulos, H., Vovk, V., and Gammerman, A. (2002b). Qualified prediction for large data sets in the case of pattern recognition. In *ICMLA*, pages 159–163.

[181] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387.

[182] Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341–356.

[183] Pedersen, J. (1978). Fiducial inference. *International Statistical Review/Revue Internationale de Statistique*, pages 147–170.

[184] Peters, G. (2014). Rough clustering utilizing the principle of indifference. *Information Sciences*, 277:358–374.

[185] Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In *International conference on machine learning*, pages 2817–2826. PMLR.

[186] Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[187] Plaut, B., Nguyen, K., and Trinh, T. (2024). Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *arXiv preprint arXiv:2402.13213*.

[188] Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, 10(37):25–42.

[189] Postels, J., Segu, M., Sun, T., Sieber, L., Van Gool, L., Yu, F., and Tombari, F. (2021). On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649*.

[190] Qu, J., Chen, Y., Yue, X., Fu, W., and Huang, Q. (2024). Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:84645–84668.

[191] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[192] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

[193] Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, page 269. NIH Public Access.

[194] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2018). Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*.

[195] Reineking, T. (2014). *Belief functions: theory and algorithms*. PhD thesis, Universität Bremen.

[196] Rogova, G. (2008). Combining the results of several neural network classifiers. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 683–692.

[197] Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. (2019). Experience replay for continual learning. *Advances in neural information processing systems*, 32.

[198] Rosenfeld, E., Ravikumar, P., and Risteski, A. (2022). An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2657. PMLR.

[199] Ross, S., Chaib-draa, B., and Pineau, J. (2007). Bayes-adaptive pomdps. *Advances in neural information processing systems*, 20.

[200] Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al. (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

[201] Rudner, T. G., Chen, Z., Teh, Y. W., and Gal, Y. (2022). Tractable function-space variational inference in Bayesian neural networks. *Advances in Neural Information Processing Systems*, 35:22686–22698.

[202] Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4):105–114.

[203] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

[204] Sale, Y., Caprio, M., and Höllermeier, E. (2023). Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*, pages 1795–1804. PMLR.

[205] Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., Zink, M., Sheffield, J., Wood, E. F., and Marx, A. (2018). Anthropogenic warming exacerbates european soil moisture droughts. *Nature Climate Change*, 8(5):421–426.

[206] Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, page 722–726, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[207] Schwarting, W., Alonso-Mora, J., and Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):187–210.

[208] Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29.

[209] Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3183–3193, Red Hook, NY, USA. Curran Associates Inc.

[210] Shafeeq, A. and Hareesha, K. (2012). Dynamic clustering of data with modified k-means algorithm. In *Proceedings of the 2012 conference on information and computer networks*, pages 221–225.

[211] Shafer, G. (1976a). *A mathematical theory of evidence*, volume 42. Princeton university press.

[212] Shafer, G. (1976b). *A mathematical theory of evidence*, volume 42. Princeton university press.

[213] Shafer, G. (1978). Two theories of probability. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1978, pages 441–465. Philosophy of Science Association.

[214] Shafer, G. (1984). The combination of evidence. Technical Report 162, School of Business, University of Kansas.

[215] Shaker, M. H. and Hüllermeier, E. (2021). Ensemble-based uncertainty quantification: Bayesian versus credal inference. In *PROCEEDINGS 31. WORKSHOP COMPUTATIONAL INTELLIGENCE*, volume 25, page 63.

[216] Singh, A., Chau, S. L., Bouabid, S., and Muandet, K. (2024). Domain generalisation via imprecise learning. *arXiv preprint arXiv:2404.04669*.

[217] Smets, P. (1993). Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of approximate reasoning*, 9(1):1–35.

[218] Smets, P. (2005a). Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147.

[219] Smets, P. (2005b). Decision making in the tbm: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147.

[220] Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial intelligence*, 66(2):191–234.

[221] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[222] Stadler, M., Charpentier, B., Geisler, S., Zügner, D., and Günnemann, S. (2021). Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048.

[223] Stein, A. L. (2020). Artificial intelligence and climate change. *Energy Engineering (Energy) eJournal*.

[224] Strat, T. M. (1984). Continuous belief functions for evidential reasoning. In *AAAI*, pages 308–313.

[225] Su, Z.-g. and Denoeux, T. (2018). Bpec: Belief-peaks evidential clustering. *IEEE Transactions on Fuzzy Systems*, 27(1):111–123.

[226] Sultana, M., Yorke-Smith, N., Wang, K., Manchingal, S. K., Mubashar, M., and Cuzzolin, F. (2025). Epistemic wrapping for uncertainty quantification. *arXiv preprint arXiv:2505.02277*.

[227] Sun, S., Zhang, G., Shi, J., and Grosse, R. (2019). Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*.

[228] Swaminathan, A. and Joachims, T. (2015). The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28.

[229] Tang, X., Yang, K., Wang, H., Wu, J., Qin, Y., Yu, W., and Cao, D. (2022). Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(4):849–862.

[230] Tao, L., Dong, M., and Xu, C. (2023). Dual focal loss for calibration. In *International Conference on Machine Learning*, pages 33833–33849. PMLR.

[231] Teeti, I., Bhargav, R. S., Singh, V., Bradley, A., Banerjee, B., and Cuzzolin, F. (2023). Temporal dino: A self-supervised video strategy to enhance action prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3281–3291.

[232] Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical science*, pages 128–139.

[233] Tong, Z., Xu, P., and Denoeux, T. (2021). An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing*, 450:275–293.

[234] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[235] Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

[236] Troffaes, M. C. and De Cooman, G. (2014). *Lower previsions*. John Wiley & Sons.

[237] Tsamoura, E., Carral, D., Malizia, E., and Urbani, J. (2021). Materializing knowledge bases via trigger graphs. *arXiv preprint arXiv:2102.02753*.

[238] Ulmer, D. T., Hardmeier, C., and Frellsen, J. (2023). Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*.

[239] Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR.

[240] Van de Ven, G. M. and Tolias, A. S. (2019). Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

[241] Van der Vaart, P., Yorke-Smith, N., and Spaan, M. T. (2024). Bayesian ensembles for exploration in deep q-learning. In *The Sixteenth Workshop on Adaptive and Learning Agents*.

[242] Vega, M. A. and Todd, M. D. (2022). A variational Bayesian neural network for structural health monitoring and cost-informed decision-making in miter gates. *Structural Health Monitoring*, 21(1):4–18.

[243] Vovk, V. (2012). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74.

[244] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

[245] Vovk, V. and Petej, I. (2012). Venn-abers predictors. *arXiv preprint arXiv:1211.0025*.

[246] Vovk, V., Shafer, G., and Nouretdinov, I. (2003). Self-calibrating probability forecasting. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

[247] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York.

[248] Wang, H., Shao, W., Sun, C., Yang, K., Cao, D., and Li, J. (2024a). A survey on an emerging safety challenge for autonomous vehicles: Safety of the intended functionality. *Engineering*, 33:17–34.

[249] Wang, K., Cuzzolin, F., Manchingal, S. K., Shariatmadar, K., Moens, D., and Hallez, H. (2024b). Credal deep ensembles for uncertainty quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[250] Wang, K., Cuzzolin, F., Shariatmadar, K., Moens, D., and Hallez, H. (2024c). Credal wrapper of model averaging for uncertainty estimation on out-of-distribution detection. *arXiv preprint arXiv:2405.15047*.

[251] Wang, K., Shariatmadar, K., Manchingal, S. K., Cuzzolin, F., Moens, D., and Hallez, H. (2024d). Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *arXiv preprint arXiv:2401.05043*.

[252] Wang, K., Shariatmadar, K., Manchingal, S. K., Cuzzolin, F., Moens, D., and Hallez, H. (2025). Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, page 107198.

[253] Wang, Y., Shi, H., Han, L., Metaxas, D., and Wang, H. (2024e). Blob: Bayesian low-rank adaptation by backpropagation for large language models. *arXiv preprint arXiv:2406.11675*.

[254] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

[255] Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*.

[256] Wu, K. and Xiao, F. (2024). A novel quantum belief entropy for uncertainty measure in complex evidence theory. *Information Sciences*, 652:119744.

[257] Wu, M. and Goodman, N. (2020). A simple framework for uncertainty in contrastive learning. *arXiv preprint arXiv:2010.02038*.

[258] Xu, L., Krzyzak, A., and Suen, C. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435.

[259] Yang, A. X., Robeyns, M., Wang, X., and Aitchison, L. (2023). Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*.

[260] Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

[261] Zaffalon, M. (2002). The Naive Credal Classifier. *Journal of Statistical Planning and Inference - J STATIST PLAN INFER*, 105:5–21.

[262] Zaffalon, M. and Fagiuoli, E. (2003). Tree-based credal networks for classification. *Reliable computing*, 9(6):487–509.

[263] Zanger, M. A., Böhmer, W., and Spaan, M. T. (2023). Diverse projection ensembles for distributional reinforcement learning. *arXiv preprint arXiv:2306.07124*.

[264] Zanger, M. A., Van der Vaart, P. R., Böhmer, W., and Spaan, M. T. (2025). Contextual similarity distillation: Ensemble uncertainties with a single model. *arXiv preprint arXiv:2503.11339*.

[265] Zhang, J., Lou, Y., Wang, J., Wu, K., Lu, K., and Jia, X. (2021). Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Internet of Things Journal*, 9(5):3443–3456.

[266] Zheng, E., Yu, Q., Li, R., Shi, P., and Haake, A. (2021). A continual learning framework for uncertainty-aware interactive image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6030–6038.

[267] Zhou, K., Martin, A., Pan, Q., and Liu, Z.-g. (2015). Median evidential c-means algorithm and its application to community detection. *Knowledge-Based Systems*, 74:69–88.

[268] Zhou, Q., Luo, H., Bossé, É., and Deng, Y. (2024). Why combining belief functions on quantum circuits? In *International Conference on Belief Functions*, pages 161–170. Springer.

[269] Zhou, Q., Tian, G., and Deng, Y. (2023). Bf-qc: Belief functions on quantum circuits. *Expert Systems with Applications*, 223:119885.

[270] Zhu, H., Tran, T. M. T., Benjumea, A., and Bradley, A. (2023). A scenario-based functional testing approach to improving dnn performance. In *2023 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, pages 199–207. IEEE.

[271] Zou, Z., Meng, X., and Karniadakis, G. E. (2025). Uncertainty quantification for noisy inputs–outputs in physics-informed neural networks and neural operators. *Computer Methods in Applied Mechanics and Engineering*, 433:117479.

A Theories of uncertainty

Uncertainty theory (UT) is an array of theories devised to encode 'second-order', 'epistemic' uncertainty, *i.e.*, uncertainty about what probabilistic process actually generates the data, can provide a principled solution to this conundrum [11, 247]. This is the situation ML is in, for we usually ignore the form of the data-generating process at hand, even accepting that it should be modelled by a probability distribution. Many (but not all) uncertainty measures amount to convex sets of distributions or 'credal sets' (*e.g.*, p-boxes) [235], while random-sets and belief functions directly assign probability values to sets of outcomes [211], modeling the fact that observations often come in the form of sets. The paramount principle in UT is to continually refine one's degree of uncertainty (measured, *e.g.*, by how wide a convex set of models is) in the light of new evidence. All uncertainty theories are equipped with operators (playing the role of Bayes' rule in classical probability) allowing one to reason with such measures (*e.g.* Dempster's combination for belief functions) [62, 220].

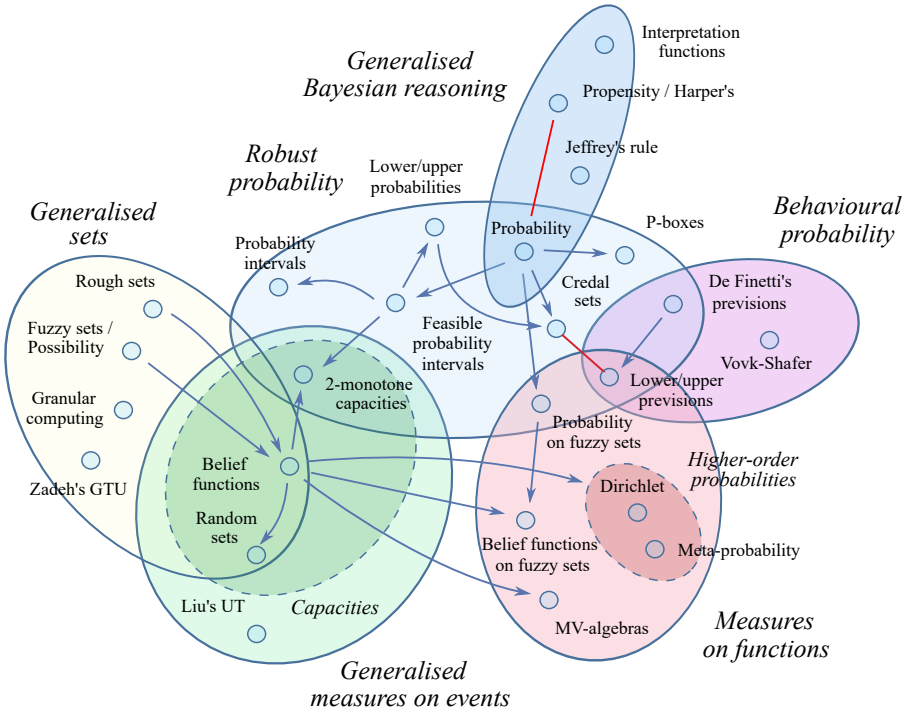


Figure 6: **Clusters of uncertainty theories.** Uncertainty theories can be arranged into various clusters based on the objects they quantify and their rationale. Arrows indicate the level of generality, with more general theories encompassing less general ones. Note that the quality and rigour of different approaches can vary significantly.

The various theories of uncertainty form 'clusters' characterised by a common rationale [51] (see Fig. 6). A first set of methods can be seen as ways of 'robustifying' classical probability: The most general such approach is Walley's theory of imprecise probability [247], a behavioural approach whose roots can be found in the ground-breaking work of de Finetti [58]. In behavioural probability, the latter is a measure of an agent's propensity to gamble on the uncertain outcomes. A different cluster of approaches hinges on generalising the very notion of set: these include, for instance, the theory of rough sets [182], possibility theory [75], and Dempster-Shafer theory [211]. More general still are frameworks generalising measure theory, *e.g.* the theory of monotone capacities or 'fuzzy measures' [95]. Some proposals (including Popper's propensity) aim at generalising Bayesian reasoning [188] in terms of either the measures used or the inference mechanisms. The theory of set-valued random variable or 'random-sets' extends the notion of set [160] and generalises Bayesian reasoning. Frameworks which completely replace events by scoring functions [244] in a functional space form arguably the most general class of methods. The diagram also illustrates the relationships between different theories, indicating which are more general and which are more specific. An arrow from formalism 1 to formalism 2 suggests that the former is a less general case of the latter.

Credal sets. In decision theory and probabilistic reasoning, *credal sets* provide a robust approach to modeling epistemic uncertainty by generalizing the traditional Bayesian framework. Unlike standard probabilistic models that assign precise probabilities to events, credal sets represent **convex sets of probability distributions**, allowing for a more flexible and cautious representation of uncertainty [247].

A *credal set* is a closed and convex set of probability distributions over a finite space Ω. This formulation allows one to express *imprecise probabilities*, where instead of a single probability value $P(A)$, we consider a range $[P^-(A), P^+(A)]$ that characterizes the lower and upper bounds of belief for an event A. This approach is particularly useful in settings where data is scarce or conflicting, making precise probability assignments unreliable [11].

Credal sets have been extensively used in *robust Bayesian inference*, *classification*, and *decision-making under ambiguity*. In machine learning, credal classifiers [261] extend Bayesian classifiers by considering sets of posterior probabilities rather than single estimates, improving robustness to small-sample uncertainties.

Moreover, *credal networks* (generalizations of Bayesian networks) allow for imprecise conditional probability tables, leading to more cautious yet reliable inferences in high-stakes applications such as medical diagnosis and risk assessment [42]. By accounting for multiple possible distributions, credal sets reduce overconfidence in decision-making. Unlike Bayesian models that rely on precise priors, credal sets allow a more agnostic approach. It is particularly useful when probability estimates come from conflicting or incomplete sources. Several credal set computation techniques are discussed in [150, 109].

Challenges in credal set computations. Credal sets, representing convex sets of probability distributions to model uncertainty, can be handled through various computational methods. One approach involves representing a credal set by its extremal points (vertices), forming a convex polytope in the probability space. These vertices can be computed using linear programming techniques [250], such as the simplex method, which navigates between vertices to find optimal solutions [57]. Alternatively, the double description method can enumerate all vertices of a convex polytope given its defining inequalities [85]. However, as the complexity of the network increases, the number of vertices can grow exponentially, leading to computational challenges. To address this, constraint-based representations define the credal set by a set of linear inequalities, offering computational efficiency, especially when the number of constraints is limited [236].

In the context of belief functions, credal sets can be derived through permutations of focal elements, but the combinatorial explosion necessitates optimization methods to manage computational load [152]. Additionally, dual representations utilize lower and upper probabilities to perform computations without explicitly considering all extreme points [249]. The choice of method depends on the specific application and the trade-off between computational efficiency and the precision required in representing uncertainty. A study [204] found that while this volume correlates with epistemic uncertainty in binary classification, its effectiveness diminishes in multi-class classification scenarios. In contrast, more recent research [111] indicates that the size of the credal set remains a reliable measure of epistemic uncertainty, even in multi-class settings, including complex datasets like ImageNet.

However, this paper does not simply advocate for credal sets, but for the adoption of second-order uncertainty measures.

B Related Epistemic AI work

Credal inference [39, 109, 204] is gradually gaining popularity as it predicts convex sets of probability distributions, known as credal sets [136], providing an alternative method for efficiently quantifying epistemic uncertainty. Credal representations [48] have been widely explored in machine learning, including the naive credal classifier [40], credal network [39], and credal random forest classification [215]. Random-sets [51] can naturally model missing data. Belief function models [49, 47], in particular, have been used for ensemble classification [142], regression [93] or to generalise max-entropy classification [50], among others.

B.1 Epistemic learning theory

Epistemic statistical learning theory, based on a 'credal' framework [31], models data-generating variability via convex sets of probabilities (credal sets) inferred from finite samples. It derives bounds

for finite hypothesis spaces (with or without realizability) and infinite model spaces, generalizing classical results. Data-dependent uniform PAC generalization bounds are also established using a random-set formulation [76].

B.2 Unsupervised learning

Unsupervised clustering is central to epistemic uncertainty research. Hard methods like c-means assign objects to single clusters, while soft methods model uncertainty, including fuzzy sets [19], possibility theory [127], rough sets [184], and evidential clustering [68, 70]. Rough sets use approximations, while evidential clustering, based on Dempster-Shafer theory [62, 211], represents uncertainty via mass functions, forming credal partitions. The ECM algorithm [157] introduced mass-based uncertainty modeling, refined by RECM [158] for dissimilarity data. BCM [141] and CCM [143] addressed meta-cluster prototype issues, while BPEC [225], MECM [267], and EGMM [114] integrated evidential reasoning. EK-NN [64], EK-NNclus [68], and EVCLUS [70] tackled clustering ambiguity, with NN-EVCLUS [67] reducing parameter dependence and enabling classification via neural networks. Key challenges include scalability, handling high-dimensional data, and ensuring robustness in uncertain environments.

B.3 Reinforcement learning

Uncertainty quantification in reinforcement learning (RL) remains challenging, with existing methods showing practical success but lacking theoretical soundness and convergence guarantees. Diverse Projection Ensembles [263] extend distributional RL by using ensemble diversity to capture epistemic uncertainty, while still modeling aleatoric uncertainty through the distribution of returns. Methods like SMC-DQN [241] combine Sequential Monte Carlo with Deep Q Networks to train model ensembles for Bayesian posterior approximation of the value function. In model-based RL, Monte Carlo Tree Search (MCTS) [25], used in AlphaZero and MuZero, is augmented with epistemic uncertainty estimates [174] to enhance strategic exploration. Research on Partially Observed Markov Decision Processes (POMDPs) [126] under epistemic uncertainty includes approaches such as Bayesian POMDPs [199] and set-valued transitions [26, 59, 60]. However, comprehensive extensions to emission probabilities and reward functions under various epistemic uncertainty types (intervals, credal/random-sets) are still lacking.

C Uncertainty estimation in uncertainty-aware models

The predictions of a classifier can be plotted in the simplex (convex hull) $\mathcal{P}$ of the one-hot probability vectors assigning probability 1 to a particular class. For instance, in a 3-class classification scenario ($\mathbf{Y} = \{a, b, c\}$), the simplex would be a 2D simplex (triangle) connecting three points, each representing one of the classes, as shown in Fig. 2 (right), which depicts all types of model predictions considered here.

C.1 Traditional Neural Networks

Traditional neural networks (NNs) predict a vector of N scores, one for each class, duly *calibrated* to a probability vector representing a (discrete, categorical) probability distribution over the list of classes $\mathbf{Y}$, $\hat{p}_{NN}(y \mid \mathbf{x}, \mathbb{D})$, which represents the probability of observing class y given the input $\mathbf{x}$ and training data $\mathbb{D}$.

C.2 Bayesian Neural Networks

Bayesian Neural Networks (BNNs) [134, 232, 92, 106] compute a predictive distribution $\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})$ by integrating over a learnt posterior distribution of model parameters θ given training data $\mathbb{D}$. This is often infeasible due to the complexity of the posterior, leading to the use of *Bayesian Model Averaging* (BMA), which approximates the predictive distribution by averaging over predictions from multiple samples. When applied to classification, BMA yields point-wise predictions.

Bayesian inference integrates over the posterior distribution $p(\theta \mid \mathbb{D})$ over model parameters θ given training data $\mathbb{D}$ to compute the predictive distribution $\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})$, reflecting updated beliefs after observing the data:

$$\hat{p}_b(y \mid \mathbf{x}, \mathbb{D}) = \int p(y \mid \mathbf{x}, \theta) p(\theta \mid \mathbb{D}) d\theta, \tag{1}$$

where $p(y \mid \mathbf{x}, \theta)$ represents the likelihood function of observing label y given x and θ. To overcome the infeasibility of this integral, direct sampling from $\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})$ using methods such as Monte-Carlo are applied to obtain a large set of sample weight vectors, $\{\theta_k, k\}$, from the posterior distribution. These sample weight vectors are then used to compute a set of possible outputs y_k, namely:

$$\hat{p}_b(y_k \mid \mathbf{x}, \mathbb{D}) = \frac{1}{|\Theta|} \sum_{\theta_k \in \Theta} \Phi_{\theta_k}(\mathbf{x}), \tag{2}$$

where Θ is the set of sampled weights, $\Phi_{\theta_k}(\mathbf{x})$ is the prediction made by the model with weights θ_k for input $\mathbf{x}$, and Φ is the function for the model. This process is called *Bayesian Model Averaging (BMA)*. BMA may inadvertently smooth out predictive distributions, diluting the inherent uncertainty present in individual models [105, 96] as shown in Fig. 7. When applied to classification, BMA yields point-wise predictions. For fair comparison and to overcome BMA's limitations, in this paper we also use sets of prediction samples obtained from the different posterior weights before averaging.

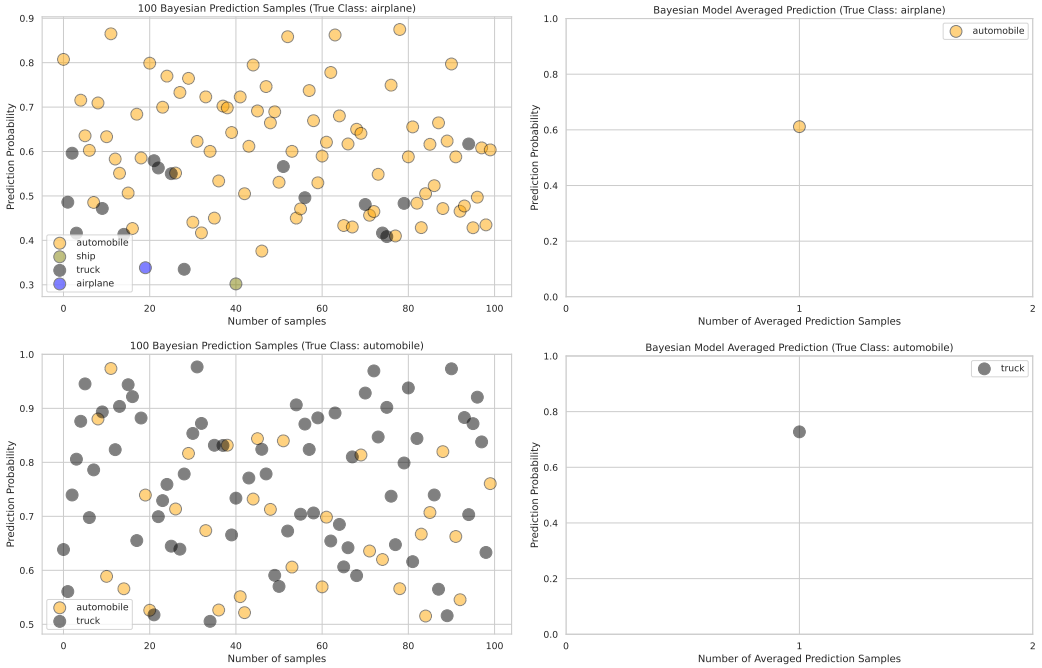


Figure 7: Visualizations of 100 prediction samples obtained prior to Bayesian Model Averaging and corresponding Bayesian Model Averaged prediction in two real scenarios from CIFAR-10.

In BNNs, *aleatoric uncertainty* is measured by the predictive entropy, while *epistemic uncertainty* is represented by *mutual information* [109], MI, which measures the difference between the entropy of the predictive distribution and the expected entropy of the individual predictions. To compute both mutual information and predictive entropy in Bayesian Neural Networks (BNNs), one utilises the predictive distributions of the model. MI quantifies the amount of information gained about the label y given the input $\mathbf{x}$ and the observed data $\mathbb{D}$, while the predictive entropy (H) measures the uncertainty associated with the predictions:

$$\text{MI}(\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})) = H(\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})) - \mathbb{E}_{\mathbb{D}}[H(p(y \mid \mathbf{x}, \theta))], \tag{3}$$

where $H(\hat{p}_b(y \mid \mathbf{x}, \mathbb{D}))$ is the entropy of the predictive distribution obtained from BMA, and $\mathbb{E}_{\mathbb{D}}[H(p(y \mid \mathbf{x}, \theta))]$ represents the expected entropy of the individual predictive distributions sampled from the posterior distribution of the parameters $p(\theta \mid \mathbb{D})$. $H(\cdot)$ denotes the Shannon entropy function.

The predictive entropy can be calculated as:

$$H(\hat{p}_b(y \mid \mathbf{x}, \mathbb{D})) = -\int \hat{p}_b(y \mid \mathbf{x}, \mathbb{D}) \log \hat{p}_b(y \mid \mathbf{x}, \mathbb{D}) dy, \tag{4}$$

where $\hat{p}(y \mid \mathbf{x}, \mathbb{D})$ is the predictive distribution. This equation represents the average uncertainty associated with the predictions across different possible values of y, considering the variability introduced by the parameter uncertainty captured in the posterior distribution $p(\theta \mid \mathbb{D})$.

C.3 Deep Ensembles

In Deep Ensembles (DEs) [132], a prediction $\hat{p}_{de}(y \mid \mathbf{x}, \mathbb{D})$ for an input $\mathbf{x}$ is obtained by averaging the predictions of K individual models: $\hat{p}_{de}(y \mid \mathbf{x}, \mathbb{D}) = \frac{1}{K} \sum_{k=1}^{K} \hat{p}_k(y \mid \mathbf{x}, \mathbb{D})$, where $\hat{p}_k$ represents the prediction of the k-th model, trained independently with different initialisations or architectures.

In Deep Ensembles, aleatoric uncertainty is assessed via the predictive entropy, averaged entropy of each ensemble's prediction, while epistemic uncertainty is encoded by the predictive variance, the difference between the entropy of all ensembles and the averaged entropy of each ensemble.

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$ denote the ensemble of K neural network models for $k = 1, 2, \ldots, K$. Given an input $\mathbf{x}$, the prediction $y_{\mathcal{M}}$ is obtained by averaging the predictions of individual models. The *predictive entropy* represents the averaged entropy of each ensemble's prediction y_k given the input $\mathbf{x}$ and the observed data $\mathbb{D}$:

$$H(\hat{p}_{de}(y \mid \mathbf{x}, \mathbb{D})) = \frac{1}{K} \sum_{k=1}^{K} H(\hat{p}_{de}(y_k \mid \mathbf{x}, \mathbb{D})), \tag{5}$$

where y_k represents the prediction of the k-th model M_k.

The *predictive variance* is measured as the difference between the entropy of all the ensembles, $H(\hat{p}_{de}(y_{\mathcal{M}} \mid \mathbf{x}, \mathbb{D}))$, and the averaged entropy of each ensemble, $H(\hat{p}_{de}(y \mid \mathbf{x}, \mathbb{D}))$.

$$H(y_{\mathcal{M}}) = H(\hat{p}_{de}(y_{\mathcal{M}} \mid \mathbf{x}, \mathbb{D})) - H(\hat{p}_{de}(y \mid \mathbf{x}, \mathbb{D})). \tag{6}$$

The predictive variance in DEs is considered an approximation of mutual information [109]. This formulation captures both the model uncertainty inherent in the ensemble predictions and the uncertainty due to the variance among individual model predictions. While DEs have proven as a good baseline method for uncertainty quantification in practice, they remain computationally expensive with several recent methods aiming to approximate ensemble uncertainties with single models [239, 130, 91, 264].

C.4 Evidential Deep Learning

Evidential Deep Learning (EDL) models [209] make predictions $\hat{p}_e(y \mid \mathbf{x}, \mathbb{D})$ as parameters of a second-order Dirichlet distribution on the class space, instead of softmax probabilities. EDL uses these parameters to obtain a pointwise prediction. Similar to BNNs, averaged DE and EDL predictions are point-wise predictions and averaging may not always be optimal.

C.5 Deep Deterministic Uncertainty

Deep Deterministic Uncertainty (DDU) [163] models differ from other uncertainty-aware baselines as they do not represent uncertainty in the prediction space, but do so in the input space by identifying whether an input sample is in-distribution (iD) or out-of-distribution (OoD). As a result, DDU provides predictions $\hat{p}_{ddu}(y \mid \mathbf{x}, \mathbb{D})$ in the form of softmax probabilities akin to traditional neural networks (NNs).

C.6 Credal Models

Models that generate *credal sets* [137, 262, 46, 9, 44] represent uncertainty in predictions by providing a set of plausible outcomes, rather than a single point estimate. A *credal set* [137, 262, 46, 9, 44] is a convex set of probability distributions on the target (class) space. Credal sets can be elicited, for instance, from predicted probability intervals [252, 29] $[\underline{\hat{p}}(y), \hat{\overline{p}}(y)]$, encoding lower and upper bounds, respectively, to the probabilities of each of the classes:

$$\hat{\mathbb{C}r}(y \mid \mathbf{x}, \mathbb{D}) = \{p \in \mathcal{P} \mid \underline{\hat{p}}(y) \leq p(y) \leq \hat{\overline{p}}(y), \forall y \in \mathbf{Y}\}. \tag{7}$$

A credal set is efficiently represented by its extremal points; their number can vary, depending on the size of the class set and the complexity of the network prediction the credal set represents.

C.7 Belief Function Models

Belief functions [212] are non-additive measures independently assigning a degree of belief to each subset A of their sample space, indicating the support for that subset.
A predicted belief function $\hat{Bel}$ on $\mathbf{Y}$ is mathematically equivalent to the credal set

$$\mathbb{C}r_{\hat{Bel}}(y \mid \mathbf{x}, \mathbb{D}) = \left\{p \in \mathcal{P} \;\middle|\; p(A) \geq \hat{Bel}(A)\right\}. \tag{8}$$

Its center of mass, termed *pignistic probability* [219] $BetP[\hat{Bel}]$, assumes the role of the predictive distribution for belief function models [233, 152]: $\hat{p}_{bel}(y \mid \mathbf{x}, \mathbb{D}) = BetP[\hat{Bel}]$.

Belief functions can be derived from *mass functions* through a normalization process, where the belief assigned to a hypothesis is the sum of the masses of all subsets of the frame of discernment that include the hypothesis. A mass function [212] is a set function [63] $m : 2^{\Theta} \to [0, 1]$ such that $m(\emptyset) = 0$ and $\sum_{A \subset \Theta} m(A) = 1$. In classification, 2^{Θ} is the set of all subsets of classes $\mathcal{C}$, the powerset $\mathbb{P}(\mathcal{C})$. Subsets of $\Theta = \mathcal{C}$ whose mass values are non-zero are called *focal elements* of m. The *belief function* associated with m is given by: $Bel(A) = \sum_{B \subseteq A} m(B)$. The redistribution of mass values back to singletons from focal sets is achieved through the concept of *pignistic probability* [219]. Pignistic probability ($BetP$), also known as Smets' pignistic transform, is a method used to assign precise probability values to individual events based on the belief function's output.

Aleatoric uncertainty in such models is represented as the pignistic entropy of predictions H_{BetP}, whereas *epistemic uncertainty* can be modelled by the 'size' of the credal set (Eq. 8).

D A comparison of uncertainty estimation models

In Tab. 1, we present the training and inference times (computational costs) for the uncertainty methods discussed in Sec. 3 and Fig. 2. Two examples of each model type are shown, all trained on the ResNet50 backbone. More training details are given below.

The models evaluated include a range of uncertainty estimation frameworks: traditional model (ResNet50), Bayesian approximations such as Laplace [106] and function-space variational inference [201], ensemble methods including deep ensembles [132] and epistemic neural networks (ENN) [176], evidential approaches [209, 190], and Epistemic AI frameworks based on credal sets [249] and random-set theory [152]. These models differ not only in their uncertainty modeling principles but also in computational costs (see Tab. 1), reflecting a spectrum of trade-offs between performance and efficiency. For instance, function-space Bayesian methods provide uncertainty but at a high computational cost [201], while epistemic random-set models offer competitive accuracy with efficient inference [152].

Table 1: Training (in minutes; per 100 epochs) and inference time (in milliseconds; per sample) comparison of uncertainty estimation methods on the CIFAR-10 dataset.

MODEL	TRAINING TIME (100 EPOCHS) (MIN)	INFERENCE TIME (MS/SAMPLE)
TRADITIONAL	85.33	1.91 ± 0.7
DETERMINISTIC (DDU) [163]	243.85	59.35 ± 0.40
BAYESIAN (LAPLACE) [106]	107.90	7.11 ± 0.89
BAYESIAN (FUNCTION SVI) [201]	1518.35	340.25 ± 0.76
ENSEMBLE (DEEP ENSEMBLES) [132]	426.66	13163.50 ± 3.37
ENSEMBLE (ENN) [176]	712.30	3.10 ± 0.03
EVIDENTIAL (EDL) [209]	188.57	6.12 ± 0.01
EVIDENTIAL (HYPER-OPINION EDL) [190]	186.56	23.01 ± 0.15
EPISTEMIC AI (CREDAL) [249]	122.95	63.0 ± 1.1
EPISTEMIC AI (RANDOM-SET) [152]	113.23	1.91 ± 0.02

Training details. All models were trained using a ResNet50 backbone (excluding the final classification layer), followed by two additional dense layers with 1024 and 512 neurons, respectively, using ReLU activation. For the *Epistemic: Random-set* [152] model, the output layer used a sigmoid activation function to support multi-label classification, while all other models, *Epistemic: Interval* [252], *Epistemic: Credal* [249], *Epistemic: Wrapper* [250], *Bayesian: Laplace* [106], *Bayesian: Function SVI* [201], *Ensemble: Deep* [132], and *Ensemble: ENN* [176], used a softmax output for multi-class classification. The initial learning rate was set to 1e-3, with a scheduler that reduced the rate by a factor of 0.1 at epochs 80, 120, 160, and 180. Models were trained for 200 epochs using

a batch size of 128. The optimizer varied by model: Adam was used for *Epistemic: Random-set*, *Epistemic: Wrapper*, *Ensemble: ENN*, and *Ensemble: Deep*, while *Bayesian: Function SVI* used SGD. Training **dataset** sizes were as follows: CIFAR-10 used 40,000 samples and ImageNet used 1,172,498 samples. Test datasets contained 10,000 samples for CIFAR-10 and 2,000 for ImageNet. For out-of-distribution (OoD) evaluation, 10,000 test samples were used. All models were trained and evaluated using 224×224 input image size with data augmentation including random horizontal/vertical shifts (magnitude 0.1) and horizontal flips. Experiments were conducted using an NVIDIA A100 80GB GPU.

The pairwise plots in Fig. 8 illustrate the relationships between key uncertainty and performance metrics: Entropy, Expected Calibration Error (ECE), Area Under the Receiver Operating Characteristic curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC), across different uncertainty estimation methods. These methods are categorized into Epistemic AI and competitor types, revealing distinct clusters and trends. Notably, Epistemic AI models show higher entropy values and competitive AUROC/AUPRC scores, indicating richer uncertainty quantification alongside robust out-of-distribution (OoD) detection. In contrast, the competitor generally exhibit lower entropy and slightly varied calibration performance. The correlations visible in the plots reflect inherent trade-offs: higher uncertainty often aligns with better OoD detection but may impact calibration, which is crucial for reliable decision-making.

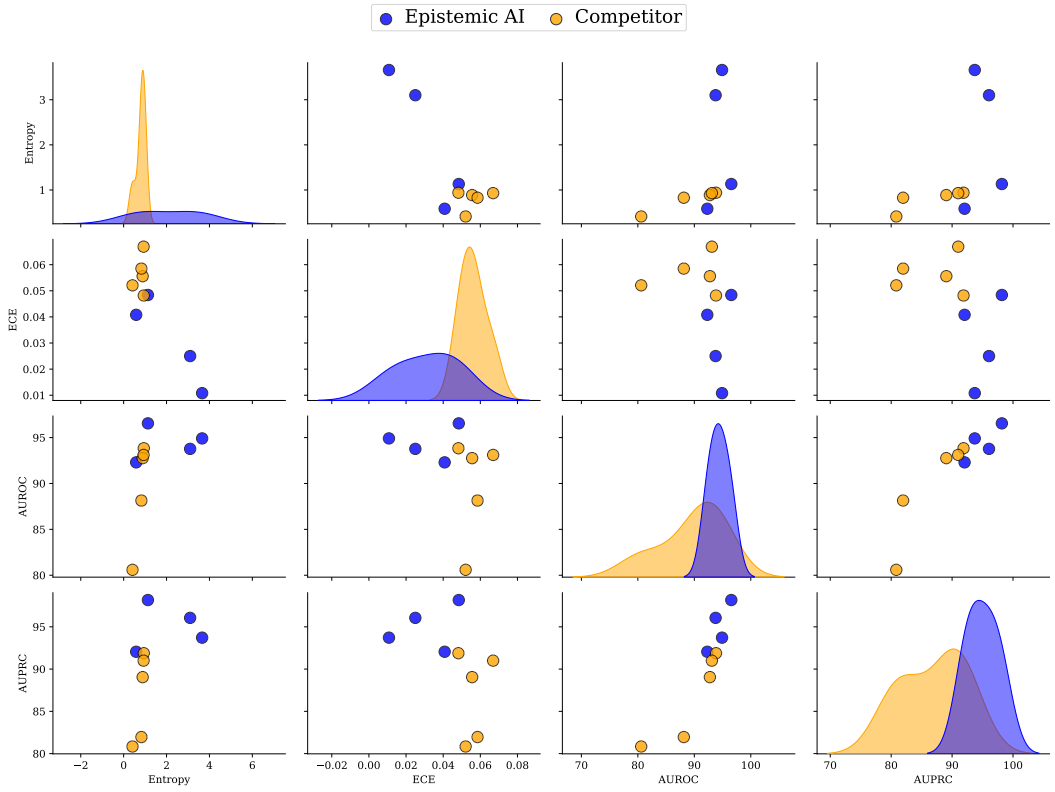


Figure 8: Pairwise comparison of uncertainty and performance metrics for various uncertainty estimation methods on CIFAR-10. Metrics include Entropy, Expected Calibration Error (ECE), AUROC, and AUPRC (OoD Detection).

Impact Statement

This work advances Epistemic AI, offering a more reliable and interpretable approach to uncertainty quantification in machine learning. By improving AI's ability to distinguish between known and unknown uncertainties, this research enhances robustness in critical applications such as healthcare, autonomous systems, climate modeling, and scientific discovery. A key ethical advantage is its

potential to mitigate overconfidence in AI predictions, reducing risks in safety-critical domains like medical diagnosis and autonomous decision-making.

Future societal impacts include more trustworthy AI systems that can adapt to novel and evolving situations, fostering responsible deployment in high-stakes environments. Furthermore, integrating epistemic uncertainty into AI could bridge gaps between symbolic reasoning and deep learning, advancing neurosymbolic AI and promoting generalizable, human-aligned decision-making. However, ethical considerations only arise in the potential misuse of uncertainty-aware AI, such as adversarial exploitation or biased decision-making if epistemic uncertainty is misinterpreted. Addressing these risks requires transparent AI models, regulatory oversight, and interdisciplinary collaboration.