Research Statement: From Statistical to Epistemic Machine Learning

Siu Lun Chau
College of Computing & Data Science, Nanyang Technological University, Singapore

Abstract—I aim to understand *how to design uncertainty-aware intelligent systems that can recognise and communicate the limits of their knowledge*. Achieving this goal requires principled approaches to modelling two fundamentally different types of uncertainty. The first, *statistical* (or aleatoric) uncertainty, arises from inherent stochasticity in the data-generating process and is typically captured using probability distributions. The second, *epistemic* uncertainty, pertains to the state of knowledge and arises from incomplete information, limited data, unjustified assumptions, or structural ambiguity. While statistical machine learning—concerned primarily with modelling aleatoric uncertainty—is a well-established and mature field, the systematic treatment of epistemic uncertainty in machine learning has received comparatively less attention. However, with the growing deployment of machine learning models in safety-critical domains such as healthcare, law, and autonomous systems, addressing epistemic uncertainty is becoming increasingly urgent. In such settings, the ability of a system to *recognise and communicate what it does not know* is just as important as making accurate predictions.

My research journey began by addressing challenges in kernel methods and Gaussian processes—two foundational tools in modern statistical machine learning. Over time, my focus has shifted toward designing algorithms that relax overly stringent or unjustified modelling assumptions, incorporate imprecise and ambiguous information, and support principled decision-making in the presence of unresolved conflict or uncertainty. To this end, I find the framework of imprecise probabilities, which generalises classical probability theory to accommodate indecision, ambiguity, and partial knowledge, both natural and powerful for building the next generation of uncertainty-aware systems. In Section 1, I will discuss my scientific achievements in the area of statistical and epistemic machine learning, followed by my vision for future research in Section 2.

1 Scientific Achievements

My research endeavour has led to several scientific contributions at the flagship conferences and journals in machine learning (NeurIPS [1–6], ICML [7, 8], UAI [9], AISTATS [10–12], AAAI [13], ECML-PKDD [14], IEEE [15], TMLR [16], Nature Communications [17]), as well as several preprints [18–22]. I was also honoured to receive first place in the IJAR Young Researcher Award 2025. In Sections 1.1 and 1.2, I outline my research contributions in statistical and epistemic machine learning, respectively.

1.1 Statistical Machine Learning with Kernels

Statistical machine learning (StatML) can be broadly understood as the study of inference and prediction algorithms that reason statistically about probability distributions, which we have access only through independent and identically distributed (i.i.d.) samples. An important challenge in this setting is to represent and manipulate distributions in a data-driven and nonparametric manner to avoid model mispecification. Kernel methods—particularly kernel mean embedding (KMEs) ([23, 24]— provide such functionality by embedding distributions into a reproducing kernel Hilbert space, which can be estimated through samples only.

Let $\mathcal{X}$ denote the input space, which may be $\mathbb{R}^d$ or more complex domains such as images, strings, or graphs, provided a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be defined. The core idea of kernel methods is to perform linear operations in a high-dimensional feature space that implicitly encode nonlinear transformations of the original inputs. This leads to highly flexible models that remain tractable with standard linear algebraic operations. In particular, due to the reproducing property of RKHSs $\mathcal{H}_k$, any $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$ satisfies $f(x) = \langle f, k(x, \cdot)\rangle$, which underpins the computational tractability and theoretical elegance of kernel methods. Furthermore, consider X some random variable in $\mathcal{X}$ with distribution $P(X)$, the kernel mean embedding

$$P(X) \mapsto \mu_P := \mathbb{E}_P[k(X, \cdot)] \in \mathcal{H}_k$$

embeds the distribution $P(X)$ into $\mathcal{H}_k$, which can be consistently estimated through empirical averages $\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)$. For a broad class of kernels, known as *characteristic kernels*, the mapping is injective—that is, $\mu_P = \mu_Q$ if and only if $P = Q$. This property provides a powerful foundation for the design of various discrepancies such as maximum mean discrepancies (MMD) [25] $\|\mu_P - \mu_Q\|_{\mathcal{H}_k}$ and Hilbert space independence criterion (HSIC) [26] $\|\mu_{P\times Q} - \mu_P \otimes \mu_Q\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.

On the other hand, kernels also play a critical role in Bayesian ML as covariance functions in Gaussian processes (GPs) [27]. GPs follow the Bayesian learning procedure: first, we place a prior $P(f)$ over functions of interest, and upon conditioning on observed data D via our specified likelihood $P(D; f)$, yield a posterior distribution $P(f \mid D)$ over functions. This allows us to model a problem at hand probabilistically, and the posterior over functions naturally captures the epistemic uncertainty during the learning procedure. For many problems, this posterior covariance is available in closed form:

$$\kappa_D(x, x') = k(x, x') - k(x, X_D)(K_{X_D X_D} + \sigma I)^{-1} k(X_D, x'),$$

leading to important breakthroughs in applications such as probabilistic numerics [28], Bayesian optimisations [29], and active learning [30]. GPs and kernel methods share deep equivalence in various cases [31]. For example, kernel ridge regressor matches the posterior mean in GP regression problems, despite being derived from entirely different principles—one rooted in a frequentist regularised loss optimisation perspective, the other in Bayesian inference. Given their deep connection, it is natural to ask: *Can we leverage GP formulation to learn distributional representations in RKHSs?* Doing so allows us to bridge the two fields further and leverage the best of both worlds—enjoy probabilistic modelling while manipulating distributional representations.

Indeed, Flaxman et al. [32] took a first step in this direction and proposed a GP-based approach to estimate KMEs μ_P, later leading to a Bayesian kernel two-sample test [33]. Nonetheless, in scenarios where modelling relationships across variables is important—such as in causal inference [34], dynamical systems [23], reinforcement learning [35]—embedding the marginal distribution as in KME does not suffice. Conditional mean embedding (CME) [36, 37] instead embed the conditional distributions $P(Y \mid X = x) \mapsto \mu_{Y|X=x} \in \mathcal{H}_\ell$ for some positive definite kernel $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. While CMEs are defined analogously to KMEs, their estimation requires a different strategy as we mostly only observe paired samples $\{(x_i, y_i)\}_{i=1}^n$. Specifically, estimating the conditioning embedding map $\hat{\mu}_{Y|X} : x \mapsto \mu_{Y|X=x}$ is equivalent to solving a vector-valued kernel ridge regression problem [38], leading to

$$\hat{\mu}_{Y|X}(x) = k(x, X)(K_{XX} + n\lambda I)^{-1}\ell(Y, \cdot) \in \mathcal{H}_\ell.$$

Research contributions to Kernel methods ∩ GPs

Bayesian CME. In [1], we proposed the first GP-based CME estimation procedure based on vector-valued Gaussian processes, leading to a *Bayesian Conditional Mean Embedding* model. By treating CME as a bi-input function $\bar{\mu}_{Y|X}(x, y) =: \mu_{Y|X=x}(y)$, we can place a GP prior $\mu_{gp} \sim \mathcal{GP}(0, k \otimes r_\ell)$ over $\bar{\mu}_{Y|X}$. However, as [39] has shown that sample paths of GPs fall outside the RKHS of their covariance kernel with probability one, extra care was needed to ensure the posterior distribution over CMEs $\mu_{gp}(x, \cdot)$ is supported in the intended RKHS $\mathcal{H}_\ell$, leading to the choice of our nuclear dominant kernel r_ℓ over ℓ [40]. This framework provides a principled Bayesian approach to learning distributional embeddings, yielding not only point estimates but also posterior covariances. In particular the posterior covariance $\kappa((x, y), (x', y')) =$

$$k_{xx'}r_{yy'} - (k_{xX}(K_{XX} + n\lambda I)^{-1}k_{Xx'})(r_y Y R_{YY}^{-1} r_{Yy'}).$$

quantifies estimation uncertainty around CMEs and proves valuable in downstream tasks. We demonstrated its effectiveness in a causal Bayesian optimisation [41] setting, where access to posterior uncertainty over CMEs enabled optimal treatment selection in a causal data fusion problem.

Deconditional GP. Besides embedding conditional distributions, kernel embeddings also allow for "reversing" conditional expectations and can solve a challenging inverse problem: *Given observations that are themselves aggregates (conditional expectations), e.g. observe* $(x_i, z_i)_{i=1}^n$ *such that* $z_i = \mathbb{E}[f(Y) \mid X = x_i]$, *can I recover* f *and ideally in a probabilistic manner?* These problems frequently occur in climate science, where statistical downscaling is required to enhance low-resolution images, and opinion pooling, where opinions are collected in a coarse manner due to cost constraints. To tackle this, we first introduced the *Conditional Mean Processes*, which allows us to model $g(x) := \mathbb{E}[f(Y) \mid X = x]$ as a GP with covariance based on the CME $\mu_{Y|X=x}$. Specifically, for $f \sim \mathcal{GP}(0, \ell)$, we have $g \sim \mathcal{GP}(0, \kappa)$ with $\kappa(x, x') = \mu_{Y|X=x}^\top \mu_{Y|X=x'}$. Denote the conditional expectations observation as $\mathbf{z}$, then we show that through inter-domain GP formulation [42], we can recover a posterior $P(f \mid \mathbf{z})$, in which the posterior mean coincides with the frequentist deconditional mean function proposed in [43]. I further showed that for the frequentist deconditional mean embedding, the learning procedure can be expressed as a two-staged vector-valued reconstruction problem and through that established a minimax optimal convergence rate under mild assumptions.

My line of work in learning distribution representations through a Bayesian manner inspired further developments in Bayesian optimisations [44], probabilistic numerics [45], sequential decision-making [46, 47], and uncertainty quantification for causal inference [48].

Research contribution to Kernel methods

Flexible models for graph and preference learning. As mentioned, the reproducing property and the representer theorem [49] make kernel methods a powerful and versatile tool for modelling a wide range of learning problems. In [15], my colleagues and I tackled the graph topology learning problem by modelling latent graph signals as functions in an RKHS. This approach relaxes the classical i.i.d. assumption required for graph signals and enables graph learning under challenging conditions such as heavy noise, missing values, and complex dependencies. In a similar vein, in [14],my colleagues and I proposed modelling latent skill vectors in ranking problems as RKHS functions, leading to a new class of nonparametric spectral ranking models. These models draw on seriation techniques[50], singular value decomposition (SVD) [51], and canonical correlation analysis [52], and crucially, support the incorporation of player covariates for *rank prediction*—a feature not available in standard spectral ranking methods.

Competitive alternative to kernel means. Besides building flexible kernel-based predictors, in [8] my colleagues and I challenged the conventional use of kernel means $\int k(X, \cdot)dP(X)$ for distributional representations. Instead, we propose to consider directional quantiles in Hilbert spaces [53], motivated by the fact that quantiles in general encapsulate more information about the distribution than the mean alone. Let $\rho^\alpha_{u\#P}u(X)$ be the α-quantile of the real-valued function distributed according to $P(u(X))$, then we define the *kernel quantile embeddings* (KQEs) as the mapping

$$P \mapsto \{\rho^\alpha_{u\#P}u(X) : \alpha \in [0, 1], u \in \mathcal{H}_k, \|u\|_{\mathcal{H}_k} = 1\}.$$

These KQEs naturally defines a family a statistical distances

Figure 1: My research in statML with kernels revolves around two distinct yet deeply connected fields: **(a)** The conditional mean embedding $\mu_{Y|X=x} \mapsto \mathbb{E}[\ell(Y, \cdot) \mid X = x]$. **(b, left)** Samples from a Gaussian process prior; **(b, right)** Samples from a Gaussian process posterior upon conditioning on observations. Larger fluctuations can be seen in unobserved regions.

that (i) are valid probability metrics under weaker conditions on the kernel than those required for KMEs, (ii) recover a kernelised analogue of the sliced Wasserstein distance, and (iii) admit efficient estimators with near-linear computational cost, in contrast to the quadratic cost of KME-based metrics. Our findings demonstrate the potential of KQEs as a powerful alternative to traditional mean-based representations.

Explainable kernel methods. While kernel methods provide flexible nonparametric models, their black-box natures hinder their use in critical applications. Inspired by the application of game theoretic concepts in feature attribution problems [54], in [3], my colleagues and I proposed the first *RKHS-SHAP* algorithm that computes Shapley value for kernel methods that provides better statistical performance than standard model-agnostic approaches. Specifically, given a cooperative game $([d], \nu)$ where $[d] = \{1, \ldots, d\}$ is the feature indices and $\nu : 2^d \to \mathbb{R}$ measures the payoff of feature subgroups, for a predictive function f and input x, the Shapley value $\phi_{x,j}(f)$ for feature j in x, which quantifies its contribution to the prediction $f(x)$, can be computed as:

$$\phi_{x,j}(f) = \sum_{S \subseteq [d] \setminus \{j\}} c_{|S|}[\nu(S \cup i) - \nu(S)],$$

for some constant terms $c_{|S|}$. Shapley values are uniquely characterised by three desirable axioms—efficiency, symmetry, and linearity—which have made them a central tool in explainable AI. Specifically, we showed that the commonly used value function $\nu(S) = \mathbb{E}[f(X) \mid X_S = x_S]$, which measures the average prediction when features in S^c are integrated out, can be efficiently estimated through CMEs, as $\nu(S) = \langle f, \mu_{X|X_S=x_S} \rangle$. This circumvents the need for density estimation and avoids model specification. In [4], my colleagues and I extended RKHS-SHAP to the preference learning setting, introducing *Pref-SHAP*, the first method to explain nonparametric preference functions using Shapley values. While RKHS-SHAP addresses statistical efficiency, in our follow-up work [20], my colleagues and I focused on improving the computational efficiency of Shapley value computation for kernel methods. By exploiting the decompositional structure of product kernels, we introduced the PKeX-Shapley (Product Kernel Exact Shapley) algorithm, which reduces the computational cost of computing exact Shapley from $O(2^d)$ to $O(d^2)$. This remarkable speed up makes our algorithms not only axiomatically driven, but also computationally tractable for industry practices. Together, RKHS-SHAP and PKeX-Shapley form a toolkit for performing interpretable, kernel-based statistical inference—bridging the gap between theoretical rigour and practical interpretability in modern machine learning.

Research contributions to GPs

GP for preference learning. I have also made substantial contributions to the field of Gaussian processes (GPs). In [10], my colleagues and I revisited the classical work of Chu and Ghahramani [55] on preference learning with GPs and critically examined their assumption that preference transitivity must be imposed. We introduced a *generalised preferential GP* model that can model general preference relationships beyond transitivity. This result establishes the universal approximation property of our model and provides a solid theoretical foundation for learning flexible, non-transitive preference structures within the GP framework.

Speeding up GP computation. GPs are often considered to be computationally heavy due to various matrix-vector multiplication procedures. To address this, in [5], my colleagues and I introduced the *Faster-Fast and Free Memory Method* (F^3M) that extends the classical classical *Fast Multipole Method* [56] to perform kernel matrix vector multiplication, a core operation for GP (and kernel methods), on a single GPU for tall and skinny data ($n \sim 10^9$, $d \leq 7$) in under a minute with user-specified error tolerance, providing enormous speed-up over existing methods.

Explainable GP and their applications. GPs, like kernel methods, are inherently black-box models due to their nonparametric nature. In [6], building on the RKHS-SHAP framework, my colleagues and I introduced the first SHAP algorithm specifically tailored for GP models, termed *GP-SHAP*. Unlike standard predictive models that return deterministic point estimates, GPs are fully probabilistic—each prediction $f(x)$ is a Gaussian random variable characterised by both a mean and a variance. This fundamental distinction requires rethinking how feature attributions should be computed in the presence of model uncertainty, leading us to formulate a stochastic cooperative game framework for probabilistic models and subsequently show how *stochastic Shapley values* can be derived and estimated in closed

form for GPs. Our recent follow-up work [22] also managed to reduce the computational complexity for computing stochastic Shapley values for a specific type of GPs, known as Functional ANOVA GP [57], from exponential to quadratic run-time, while retaining the ability to propagate predictive uncertainty into explanation uncertainty.

In [11], my colleagues and I integrated *GP-SHAP* into Bayesian optimisation (BO) and demonstrated how, in conjunction with a preference learning module, it enables the design of a collaborative (via preference learning) and explainable (via GP-SHAP) BO procedure. Beyond enabling individual experts to express their preferences within a BO procedure, in [21], my colleagues and I studied a *group preferential* BO problem, where the goal is to use BO to facilitate consensus among a group of experts in a resource-efficient manner. To demonstrate the real-world impact of these methodologies beyond the ML community, in [17], my colleagues and I applied BO to high-throughput chemical reaction optimisation and achieved state-of-the-art performance, translating cutting-edge modelling techniques into practical scientific discovery.

1.2 Epistemic Machine Learning with Imprecise Probabilities

More recently, my research has focused more on epistemic machine learning (EpiML), which broadly speaking, investigates how a rational agent should make predictions and decisions under epistemic uncertainty. Unlike statistical (aleatoric) uncertainty, which arises from inherent randomness in distributions, epistemic uncertainty stems from a lack of knowledge. This may be due to limited data, but also includes deeper sources such as potential distribution shifts, unknown strategic behaviour, imprecise or misspecified hypothesis spaces, and indeterminate or conflicting side information. Addressing these challenges requires representations that can meaningfully capture ignorance, imprecision, and indeterminacy. However, as numerous scholars have argued [58–60], such epistemic uncertainty, rooted in "not-knowing" rather than randomness, cannot be faithfully modelled using precise probability measures alone. This calls for a shift toward alternative uncertainty frameworks, such as imprecise probabilities, to support cautious yet principled reasoning under epistemic limitations.

My recent work (see Figure 2) in this direction has centred on incorporating explicit representations of imprecision into several foundational aspects of statistical machine learning, including: (1) prediction, (2) hypothesis testing, (3) belief elicitation, and (4) the design of probability metrics. The overarching goal of these efforts is to demonstrate that not only is it more natural and principled to reason and learn while acknowledging epistemic uncertainty, but doing so can also lead to practically effective outcomes. Even when we explicitly acknowledge imprecision, we can still achieve meaningful performance, showing that the world does not collapse if we choose to be a bit more cautious, uncertainty-aware, and epistemically humble.

Learning under potential distribution shifts

In [7], colleagues and I revisited the domain generalisation (DG) problem through the lens of imprecise probability. DG [61] concerns learning a model $f : \mathcal{X} \to \mathcal{Y}$ from multiple datasets drawn i.i.d. from distributions $P_1, \ldots, P_m$, such that it performs reliably on an unseen but related target distribution $P_\star$. Our contributions are threefold. First, we clarify and disentangle two key sources of uncertainty in DG: aleatoric uncertainty, due to limited data within each domain, and epistemic uncertainty, due to ignorance about $P_\star$. Second, we reinterpret the widely used convex hull $\mathrm{ConvHull}(P_1, \ldots, P_m)$—typically motivated by computational convenience [62, 63]—as a finitely generated credal set, thus giving DG problem a principled behavioural foundation rooted in IP. Third, we challenge the conventional approach of resolving ambiguity prior to learning, i.e., by selecting some $P' \in C$ for risk minimisation,

$$f' = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{P'}[L(f(X), Y)]$$

for some lost function L. Instead, we argue that such ambiguity should be preserved during training and only resolved at deployment, based on the practitioner's preferences, rather than prematurely by the model developer during training.

To operationalise this, we introduce a new model class: augmented hypothesis functions $h\colon \mathcal{X} \times \Theta \to \mathcal{Y}$, which allow post-training selection of $\theta \in \Theta$ to reflect different preferences, effectively navigating the credal set C. Selecting a particular θ yields a predictive model $h(\cdot, \theta)$ as if one had first chosen θ (corresponding to a specific $P' \in C$) and then trained a model accordingly. This corresponds to instead

$$h(\cdot, \theta) = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{P[\theta]}[L(f(X), Y)].$$

where for each θ, $P[\theta] \in C$. Our proposed learning algorithm, *Imprecise Risk Optimisation* (IRO), extends classical Multiple Gradient Descent Algorithms (MGDA) to accommodate infinitely many objectives—one for each distribution in C. Under mild regularity conditions, we prove that the learned model $h(x, \theta)$ is risk-optimal with respect to the distribution implicitly selected by θ, as if the ambiguity had been resolved prior to training. This work demonstrates how principles from IP can give clarity to ML problems and inform practical algorithms that respect underlying epistemic uncertainty.

Statistical testing under imprecise hypothesis

In [12], my colleagues and I addressed a longstanding open problem in imprecise probability: given two sources of ambiguity—each represented by a finitely generated credal set constructed from empirical samples at its extreme points—how can we statistically compare these credal sets, and in what way?

While seminal work by Huber and Strassen [64] and subsequent research [65, 66] extended the Neyman–Pearson framework to composite hypotheses of the form $H_0 : C_0$ is true versus $H_1 : C_1$ is true, our work instead seeks to generalise null hypothesis testing to explicitly accounting for the epistemic imprecision, enabling statistical comparison between credal sets derived from data.

Classical two-sample hypothesis testing typically addresses the question of whether two distributions are equal, formalised

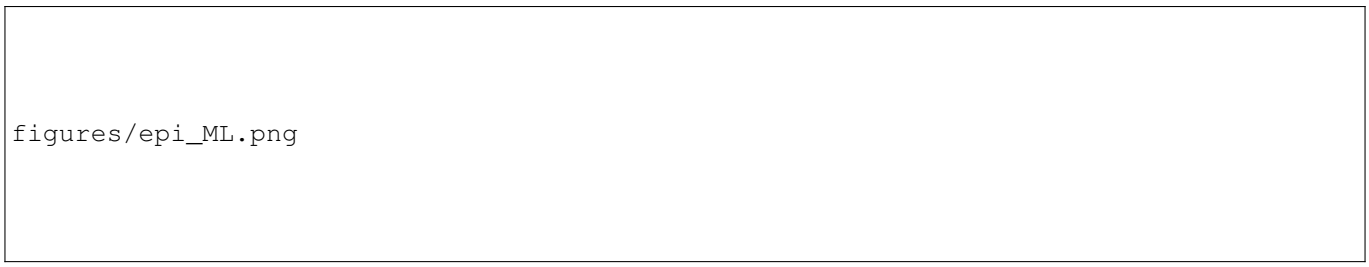

Figure 2: My projects in EpiML, which touch upon foundational aspects of machine learning. **(a) Supervised learning under potential distribution shifts:** Multiple distributions are observed at training time, how can we leverage and carry over such imprecision when designing our model h? **(b) Statistical testing under imprecise hypothesis:** Classical null hypothesis testing concerns whether P_X=P_Y or not, but when ambiguity arises in observations, we need to instead reason with imprecise hypotheses and compare credal sets instead. **(c) Truthful elicitation under indeterminate forecasters:** Proper scoring rules for precise forecaster holding P as belief are well-studied, but what if our forecaster is indeterminate? Can there still be a proper elicitation procedure? **(d) Metrics for imprecise probabilities:** Probability metrics are well studied, but what if our probability assessments are imprecisely specified? How to quantify that?

as $H_0 : P_X = P_Y$ versus $H_1 : P_X \neq P_Y$. When reasoning with sets of distributions, such as credal sets, there are richer notions of comparison, each lending itself to distinct statistical decision-making tasks. Let $C_X := \text{ConvHull}(P_X^{(1)}, \ldots, P_X^{(\ell)})$ and $C_Y := \text{ConvHull}(P_Y^{(1)}, \ldots, P_Y^{(r)})$ denote the credal sets of interests. Utilising samples drawn from their extreme points, we propose four types of hypotheses: 1) **specification** $H_{0,\in} : P_X \in C_Y$, 2) **inclusion** $H_{0,\subseteq} : C_X \subseteq C_Y$, 3) **equality** $H_{0,=} : C_X = C_Y$, and 4) **plausibility** $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$. Specification test can be used for credal set calibration problems [67], and the inclusion test can be used for uncertainty comparison. Equality test can be used to detect any differences in epistemic uncertainty after, e.g., a medical treatment, and plausibility test provides a distributionally robust two-sample testing procedure.

Our contribution goes beyond proposing four meaningful hypotheses for comparing credal sets—we also develop a practical, non-parametric, statistically valid testing procedure to test them. Building on advances in kernel two-sample testing [68], a powerful class of non-parametric methods, we adapt these techniques to the imprecise probabilistic setting to enable the comparison of epistemic uncertainty. Crucially, kernel methods naturally accommodate structured data types such as images and sequences, making our testing procedures broadly applicable beyond conventional tabular data. For each test, we establish asymptotic Type I error control, ensuring that the false discovery rate remains below the chosen significance level. Moreover, we prove that each test achieves asymptotic consistency, attaining zero Type II error against any fixed alternative hypothesis in the limit.

Truthful elicitation under indeterminate forecasters

Building on our work in imprecise prediction and hypothesis testing, in [9], my colleagues and I address the problem of truthful information elicitation—a foundational challenge in mechanism design with growing importance in machine learning, particularly through its connection to proper scoring mechanisms [69]. We ask: *How can we design scoring mechanisms that incentivise agents who face ambiguity to truthfully report their imprecise forecasts, for example, in the form of a credal set of probabilistic predictions?* This problem is particularly challenging due to a series of impossibility results [70, 71], which demonstrate that, unlike in the precise case, no deterministic scoring rule can strictly properly elicit imprecise beliefs. Nonetheless, we show that by explicitly linking elicitation to downstream decision-making, through the actions of a decision-maker responding to the forecast, truthful elicitation becomes achievable. Specifically, we introduce a new class of randomised strictly proper scoring rules for imprecise forecasters, which provably overcome existing impossibility results and ensure truthful reporting of credal sets.

Metrics for imprecise probabilities

In [18], my colleagues and I studied the problem of developing metrics for IP models. Probability metrics are central to ML tasks like prediction and generation, as they are typically framed as minimising discrepancies between target and model distributions. Various classes of probability metrics have been widely studied, such as ϕ-divergences [72, 73] and Bregman divergences [74, 75], their generalisation to IP models has been quite limited. Nonetheless, a well-defined and computationally efficient imprecise probability metric could lead to a principled approach to incorporating ambiguity and epistemic uncertainty into a broad range of ML problems, enabling models that reason more robustly under imprecision.

Motivated by this, my colleagues and I focused on the broad class of Integral Probability Metrics (IPMs) [76] and showed that the Choquet integral [77] enables their extension from probability measures to capacities—one of the most general classes of IP models. We introduced the resulting family of metrics as the *Integral Imprecise Probability Metrics* (IIPM), defined as

$$\text{IIPM}_{\mathcal{F}}(\nu, \mu) := \sup_{f \in \mathcal{F}} \left\{ \left| \oint f d\nu - \oint f d\mu \right| \right\},$$

where ν, μ are two capacities, $\mathcal{F}$ some function class, and $\oint$ denotes the Choquet integral [78]. Conditions on $\mathcal{F}$ under which

$\text{IIPM}_{\mathcal{F}}$ metrises the weak convergence of capacities in the sense of [79] were also investigated.

To illustrate the flexibility of the framework, we instantiated it with lower probabilities and demonstrated how different choices of $\mathcal{F}$ yield novel metrics for IP models. For instance, taking $\mathcal{F}$ as the class of bounded Lipschitz functions recovers a lower Dudley metric, while choosing indicator functions over measurable events gives a lower total variation distance. Besides comparing distinct uncertainty models, we showed that by measuring the discrepancies between an IP model with its conjugate yields a new class of epistemic uncertainty quantifiers, which we call *Maximum Mean Imprecision (MMI)*, illustrating with a lower probability,

$$\text{MMI}_{\mathcal{F}}(\underline{P}) := \text{IIPM}_{\mathcal{F}}(\overline{P}, \underline{P}).$$

We also proved that $\text{MMI}_{\mathcal{F}}$ satisfies core desiderata for uncertainty quantification [80–83]. Furthermore, we showed that for the lower total variation function class $\mathcal{F}_{TV}$, $\text{MMI}_{\mathcal{F}_{TV}}$ admits a tight linear-time computable upper bound. In experiments on selective classification [84], we demonstrated that our approach achieves comparable performance to the Generalised Hartley measure, while requiring only linear-time computation, instead of exponential time. This makes our method scalable to high-cardinality classification tasks such as CIFAR-100 [85], enabling efficient quantification of credal uncertainty in large-scale settings.

2 Vision for future research

Despite remarkable advances in AI, particularly in generative models and large language models, a critical gap remains in the ability of current systems to effectively handle uncertainty and generalise beyond their training data. One plausible path toward building ever more powerful intelligent systems is to continue investing in computational efficiency and scale—training increasingly large models on ever-expanding datasets. While this direction has yielded impressive results and substantial practical benefits, many argue that such models primarily excel at memorisation, rather than genuine understanding or reasoning.

An alternative—and, in my view, more natural—route is to fundamentally rethink how we design learning architectures and theories, with a focus on enabling models to be honest about what they do not know. This means developing AI systems that can recognise the limits of their knowledge, explicitly represent uncertainty, and make decisions accordingly. Such epistemically aware models offer a principled foundation for building AI that is not only powerful but also reliable, interpretable, and aligned with the complexities of the real world.

To advance this goal, there are a few research questions I would love to tackle in the coming years:

1. Where does epistemic uncertainty come from, and how can it be validated? Information cannot be created from nothing—and the same applies to epistemic uncertainty. In some problems, such as learning under multiple distributions or conflicting objectives, epistemic uncertainty naturally arises from structural ambiguity. However, in standard supervised learning—where a model is trained on a given dataset—it may appear that epistemic variability is introduced arbitrarily. In Bayesian machine learning, uncertainty stems from a prior over model parameters, which is updated via Bayes' rule to yield a posterior that reflects parameter uncertainty given the data. In ensemble methods, epistemic variation arises from differences in model initialisation, training subsets, or architectural choices. Some approaches [86] introduce selection criteria to discard implausible models from the ensemble. Evidential methods, by contrast, attempt to learn second-order uncertainty through optimisation—but as shown by Bengs and Waegeman [87], such methods lack axiomatic justification.

These diverse approaches all attempt to encode epistemic uncertainty—but how can we assess their quality? Which sources of variation are most meaningful for decision-making? Do they truly enable systems to report what they do not know? A careful analysis of these questions may help clarify and unify the various mechanisms currently used to make machine learning models epistemically aware.

2. How to make decisions under irreducible epistemic uncertainty? Epistemic uncertainty is often understood as uncertainty that could, in principle, be reduced through the acquisition of additional information. However, I argue that certain forms of uncertainty are fundamentally irreducible, even in the limit of unlimited "data." One example is the generalisation problem in machine learning: no matter how much data we collect from one distribution, it may provide no insight into another, particularly if that other distribution governs the deployment environment. While one might suggest gathering data from the target distribution, this may be infeasible or ill-defined, especially in predictive tasks involving the future, such as in causal inference, performative prediction, or under strategic manipulation, where present decisions can shape future data-generating processes.

Irreducible uncertainty also arises in multi-agent settings with conflicting preferences. Arrow's impossibility theorem [88] shows that no aggregation rule can reconcile all stakeholders' utilities while satisfying basic fairness axioms. Here, uncertainty does not stem from data scarcity, but from fundamental limits on justification and aggregation. The key question, then, is not how to eliminate uncertainty, but what principles or rationality axioms should govern decisions—such as treatment assignment, contract design, or model learning—when uncertainty is unavoidable. This is why I find cooperative game theory and mechanism design particularly compelling: they seek principled decisions under irreconcilable constraints. In my view, such axiomatic approaches offer a robust foundation for reasoning and acting under structural ambiguity.

3. How to make imprecise probabilities computationally efficient for ML problems? Imprecise probabilities (IP) offer philosophically grounded and mathematically rich frameworks for modelling epistemic uncertainty. However, operationalising these frameworks in ML poses significant computational challenges. For instance, representing a belief function over a discrete space of 1000 labels—as in ImageNet—would require specifying up to 2^{1000} mass assignments, which is clearly

infeasible. Moreover, many standard tools in probabilistic modelling—such as Monte Carlo and Markov chain Monte Carlo methods—remain undefined or underdeveloped in the context of IP. These techniques have long served as a bridge between abstract probability theory and practical applications in machine learning. Their absence in the IP setting is a major bottleneck—but also an opportunity. Advancing computational methods for IP will not only equip machine learning with more expressive and principled tools for handling imprecision but also contribute deeply to the mathematical foundations of uncertainty. Bridging this gap promises mutual enrichment for both communities.

Acknowledgements. I am in debt to all my collaborators and mentors; without them, my research journey would not have been possible. I extend a special thanks to Michele Caprio for his insightful feedback on the original draft.

References

[1] **Siu Lun Chau**, Jean-Francois Ton, Javier González, Yee Teh, and Dino Sejdinovic. Bayesimp: Uncertainty quantification for causal data fusion. *Advances in Neural Information Processing Systems*, 34:3466–3477, 2021.

[2] **Siu Lun Chau**, Shahine Bouabid, and Dino Sejdinovic. Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems*, 34: 17813–17825, 2021.

[3] **Siu Lun Chau**, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 35: 13050–13063, 2022.

[4] Robert Hu, **Siu Lun Chau**, Jaime Ferrando Huertas, and Dino Sejdinovic. Explaining preferences with shapley values. *Advances in Neural Information Processing Systems*, 35:27664–27677, 2022.

[5] Robert Hu, **Siu Lun Chau**, Dino Sejdinovic, and Joan Glaunès. Giga-scale kernel matrix-vector multiplication on gpu. *Advances in Neural Information Processing Systems*, 35:9045–9057, 2022.

[6] **Siu Lun Chau**, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36:50769–50795, 2023.

[7] Anurag Singh, **Siu Lun Chau**, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 45544–45570, 2024.

[8] Masha Naslidnyk, **Siu Lun Chau**, François-Xavier Briol, and Krikamol Muandet. Kernel quantile embeddings and associated probability metrics. *To Appear in ICML 2025*.

[9] Anurag Singh, **Siu Lun Chau**, and Krikamol Muandet. Truthful elicitation of imprecise forecasts. *To Appear in 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[10] **Siu Lun Chau**, Javier Gonzalez, and Dino Sejdinovic. Learning Inconsistent Preferences with Gaussian Processes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 2266–2281. PMLR, May 2022. ISSN: 2640-3498.

[11] Masaki Adachi, Brady Planden, David Howey, Michael A Osborne, Sebastian Orbell, Natalia Ares, Krikamol Muandet, and **Siu Lun Chau**. Looping in the human: Collaborative and explainable bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 505–513. PMLR, 2024.

[12] **Siu Lun Chau**, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal two-sample

tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2025.

[13] Kiet QH Vo, Muneeb Aadil, **Siu Lun Chau**, and Krikamol Muandet. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15411–15419, 2024.

[14] **Siu Lun Chau**, Mihai Cucuringu, and Dino Sejdinovic. Spectral ranking with covariates. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 70–86. Springer, 2022.

[15] Xingyue Pu, **Siu Lun Chau**, Xiaowen Dong, and Dino Sejdinovic. Kernel-based graph learning from smooth signals: A functional viewpoint. *IEEE Transactions on Signal and Information Processing over Networks*, 7:192–207, 2021.

[16] Simon Föll, Alina Dubatovka, Eugen Ernst, **Siu Lun Chau**, Martin Maritsch, Patrik Okanovic, Gudrun Thaeter, Joachim Buhmann, Felix Wortmann, and Krikamol Muandet. Gated domain units for multi-source domain generalization. *Transactions on Machine Learning Research*, 2023.

[17] Joshua W Sin, **Siu Lun Chau**, Ryan P Burwood, Kurt Püntener, Raphael Bigler, and Philippe Schwaller. Highly parallel optimisation of nickel-catalysed suzuki reactions through automation and machine intelligence. *To appear in Nature Communications*, 2025.

[18] **Siu Lun Chau**, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. *arXiv preprint arXiv:2505.16156*, 2025.

[19] Kiet QH Vo, **Siu Lun Chau**, Masahiro Kato, Yixin Wang, and Krikamol Muandet. Explanation design in strategic learning: Sufficient explanations that induce non-harmful responses. *arXiv preprint arXiv:2502.04058*, 2025.

[20] Majid Mohammadi, **Siu Lun Chau**, and Krikamol Muandet. Computing exact shapley values in polynomial time for product-kernel methods. *arXiv preprint arXiv:2505.16516*, 2025.

[21] Masaki Adachi, **Siu Lun Chau**, Wenjie Xu, Anurag Singh, Michael A Osborne, and Krikamol Muandet. Bayesian optimization for building social-influence-free consensus. *arXiv preprint arXiv:2502.07166*, 2025.

[22] Majid Mohammadi, Krikamol Muandet, Ilaria Tiddi, Annette Ten Teije, and **Chau, Siu Lun**. Exact shapley attributions in quadratic-time for fanova gaussian processes. *arXiv preprint arXiv:2508.14499*, 2025.

[23] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

[24] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[25] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19:513–520, 2006.

[26] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

[27] Carl Edward Rasmussen and Christopher K Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[28] François-Xavier Briol, Chris J Oates, Mark Girolami, Michael A Osborne, and Dino Sejdinovic. Probabilistic integration. *Statistical Science*, 34(1):1–22, 2019.

[29] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[30] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[31] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

[32] Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Association for Computing Machinery, 2016.

[33] Qinyi Zhang, Veit Wild, Sarah Filippi, Seth Flaxman, and Dino Sejdinovic. Bayesian kernel two-sample testing. *Journal of Computational and Graphical Statistics*, 31(4): 1164–1176, 2022.

[34] D. Sejdinovic. An overview of causal inference using kernel embeddings. *arXiv:2410.22754*, 2024.

[35] Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. In *Forty-first International Conference on Machine Learning*.

[36] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[37] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. ICML'12, page 1803–1810, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

[39] Milan Lukić and Jay Beder. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10): 3945–3969, 2001.

[40] Natesh S Pillai, Qiang Wu, Feng Liang, Sayan Mukherjee, and Robert L Wolpert. Characterizing the function space for bayesian kernel models. *Journal of Machine Learning Research*, 8(8), 2007.

[41] Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

[42] Miguel Lázaro-Gredilla and Anibal Figueiras-Vidal. Inter-domain gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22, 2009.

[43] Kelvin Hsu and Fabio Ramos. Bayesian deconditional kernel mean embeddings. In *International Conference on Machine Learning*, pages 2830–2838. PMLR, 2019.

[44] Mengyan Zhang, Shahine Bouabid, Cheng Soon Ong, Seth Flaxman, and Dino Sejdinovic. Indirect query bayesian optimization with integrated feedback. *arXiv preprint arXiv:2412.13559*, 2024.

[45] Zonghao Chen, Masha Naslidnyk, Arthur Gretton, and Francois-Xavier Briol. Conditional bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 648–684. PMLR, 2024.

[46] Diego Martinez-Taboada and Dino Sejdinovic. Bayesian counterfactual mean embeddings and off-policy evaluation. *arXiv preprint arXiv:2211.01518*, 2022.

[47] Diego Martinez-Taboada and Dino Sejdinovic. Sequential decision making on unmatched data using bayesian kernel embeddings. *arXiv preprint arXiv:2210.13692*, 2022.

[48] Hugh Dance, Peter Orbanz, and Arthur Gretton. Spectral representations for accurate causal uncertainty quantification with gaussian processes. *arXiv preprint arXiv:2410.14483*, 2024.

[49] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68 (3):337–404, 1950.

[50] Fajwel Fogel, Alexandre d'Aspremont, and Milan Vojnovic. Spectral ranking using seriation. *Journal of Machine Learning Research*, 17(88):1–45, 2016.

[51] Alexandre d'Aspremont, Mihai Cucuringu, and Hemant Tyagi. Ranking and synchronization from pairwise measurements via svd. *Journal of Machine Learning Research*, 22(19):1–63, 2021.

[52] David Weenink. Canonical correlation analysis. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 25, pages 81–99. University of Amsterdam Amsterdam, 2003.

[53] Ricardo Fraiman and Beatriz Pateiro-López. Quantiles for finite and infinite dimensional data. *Journal of Multivariate Analysis*, 108:1–14, 2012.

[54] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[55] Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.

[56] J Carrier, Leslie Greengard, and Vladimir Rokhlin. A fast adaptive multipole algorithm for particle simulations. *SIAM journal on scientific and statistical computing*, 9(4): 669–686, 1988.

[57] Xiaoyu Lu, Alexis Boukouvalas, and James Hensman. Additive gaussian processes revisited. In *International conference on machine learning*, pages 14358–14383. PMLR, 2022.

[58] Isaiah Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Princeton University Press, Princeton, NJ, 1980.

[59] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[60] Matthias CM Troffaes and Gert De Cooman. *Lower previsions*. John Wiley & Sons, 2014.

[61] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, 2013.

[62] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 367–374. AUAI Press, 2009.

[63] Simon Föll, Alina Dubatovka, Eugen Ernst, **Siu Lun Chau**, Martin Maritsch, Patrik Okanovic, Gudrun Thaeter,

Joachim M Buhmann, Felix Wortmann, and Krikamol Muandet. Gated domain units for multi-source domain generalization. *Transactions on Machine Learning Research*, 2023.

[64] Peter J Huber and Volker Strassen. Minimax tests and the neyman-pearson lemma for capacities. *The Annals of Statistics*, pages 251–263, 1973.

[65] Andreas Buja. On the huber-strassen theorem. *Probability Theory and Related Fields*, 73(1):149–152, 1986.

[66] Robert Hafner. Konstruktion robuster teststatistiken. *Data Analysis and Statistical Inference, Eul, Bergisch Gladbach*, pages 145–160, 1992.

[67] Mira Jürgens, Thomas Mortier, Eyke Hüllermeier, Viktor Bengs, and Willem Waegeman. A calibration test for evaluating set-based epistemic uncertainty representations. *arXiv preprint arXiv:2502.16299*, 2025.

[68] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[69] Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. *arXiv preprint arXiv:2504.01781*, 2025.

[70] Conor Mayo-Wilson and Gregory Wheeler. Accuracy and imprecision: A mildly immodest proposal. *Philosophy and Phenomenological Research*, 2015.

[71] Miriam Schoenfield. The accuracy and rationality of imprecise credences. *Noûs*, 51(4):667–685, 2017.

[72] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[73] Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[74] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7 (3):200–217, 1967.

[75] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[76] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

[77] Gustave Choquet. Theory of capacities. In *Annales de l'institut Fourier*, volume 5, pages 131–295, 1954.

[78] G. Choquet. Théorie des capacités. *Ann. Inst. Fourier 5 (1953/1954) 131–292.*, 1953.

[79] Ding Feng and Hung T Nguyen. Choquet weak convergence of capacity functionals of random sets. *Information Sciences*, 177(16):3239–3250, 2007.

[80] Joaquín Abellán and George J Klir. Additivity of uncertainty measures on credal sets. *International Journal of General Systems*, 34(6):691–713, 2005.

[81] Radim Jiroušek and Prakash P Shenoy. A new definition of entropy of belief functions in the dempster–shafer theory. *International Journal of Approximate Reasoning*, 92:49–65, 2018.

[82] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, August 2022. URL https://proceedings.mlr.press/v180/hullermeier22a.html. ISSN: 2640-3498.

[83] Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the Volume of a Credal Set a Good Measure for Epistemic Uncertainty?, June 2023. URL http://arxiv.org/abs/2306.09586. arXiv:2306.09586 [cs, stat].

[84] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and Epistemic Uncertainty with Random Forests, January 2020. URL http://arxiv.org/abs/2001.00893. arXiv:2001.00893 [cs, stat].

[85] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[86] Timo Löhr, Paul Hofman, Felix Mohr, and Eyke Hüllermeier. Credal prediction based on relative likelihood. *arXiv preprint arXiv:2505.22332*, 2025.

[87] Viktor Bengs and Willem Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation.

[88] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.