Truthful Elicitation of Imprecise Forecasts

Anurag Singh[1] **Siu Lun Chau**[2] **Krikamol Muandet**[1]

[1]Rational Intelligence Lab, CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2]College of Computing & Data Science, Nanyang Technological University, Singapore

Abstract

The quality of probabilistic forecasts is crucial for decision-making under uncertainty. While proper scoring rules incentivize truthful reporting of precise forecasts, they fall short when forecasters face epistemic uncertainty about their beliefs, limiting their use in safety-critical domains where decision-makers (DMs) prioritize proper uncertainty management. To address this, we propose a framework for scoring *imprecise forecasts*—forecasts given as a set of beliefs. Despite existing impossibility results for deterministic scoring rules, we enable truthful elicitation by drawing connection to social choice theory and introducing a two-way communication framework where DMs first share their aggregation rules (e.g., averaging or min-max) used in downstream decisions for resolving forecast ambiguity. This, in turn, helps forecasters resolve indecision during elicitation. We further show that truthful elicitation of imprecise forecasts is achievable using proper scoring rules randomized over the aggregation procedure. Our approach allows DM to elicit and integrate the forecaster's epistemic uncertainty into their decision-making process, thus improving credibility.

1 INTRODUCTION

Probabilistic forecasting is a powerful tool for decision-making under uncertainty with diverse applications ranging from energy demand forecasting [Pinson and Girard, 2012, Pinson, 2013] and credit risk assessment [Rindt et al., 2022, Yanagisawa, 2023] to machine learning (ML) [Singh et al., 2023] and large language models (LLMs) [Shao et al., 2024, Wu and Hartline, 2024]. Proper scoring rules serve as fundamental tools for evaluating the quality of probabilistic forecasts [Brier, 1950, Murphy and Winkler, 1988, Gneiting and Raftery, 2007]. They also serve as a backbone for eliciting other distributional properties such as their moments [Frongillo and Kash, 2014]. By assigning numerical scores based on the reported forecast and the realized outcome, these rules incentivize truthful reporting, i.e., any deviation from the forecaster's true beliefs would result in suboptimal scores. Beyond applications in statistics, proper scoring rules have a deep connection with mechanism design, a sub-field of economics. When used as a payment mechanism, the agents have no incentive to lie, a property known as incentive compatibility [Myerson, 1981].

Traditionally, scoring rules operate under the assumption that forecasters possess a *precise* probabilistic belief about some uncertain event. They are designed to reward the forecasters whose forecasts reflect their true precise beliefs [Savage, 1971, Gneiting and Katzfuss, 2014]. For example, in weather forecasting [Brier, 1950] a forecaster who believes there is a 60% chance of rain tomorrow should ideally report 60% as their forecast. However, in many real-world scenarios, forecasters face significant ambiguity due to the inherent complexity of atmospheric systems, coupled with limited data and model resolution, which introduce substantial imprecision [Wilks, 2011]. It is thus plausible for forecasters to report imprecise probability assessments in these scenarios; for example, the chance of rain tomorrow may be assessed within the interval $[50\%, 70\%]$. Importantly, classical proper scoring rules built for precise forecasts cannot account for such additional uncertainty [Konek, 2015].

Under the context of machine learning, imprecise forecasting is closely related to the concept of out-of-distribution (OOD) generalization [Muandet et al., 2013, Zhou et al., 2023]. In standard supervised learning, where training and test data are assumed to be independent and identically distributed (i.i.d.), the predictive model reflects the learner's precise belief about the data generating process. However, in OOD generalization—where multiple training datasets are observed, and the test data may not be i.i.d. with the training data—Singh et al. [2024] argue that the notion of generalization (e.g., average-case or worst-case optimization strategy)

Accepted for the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025).

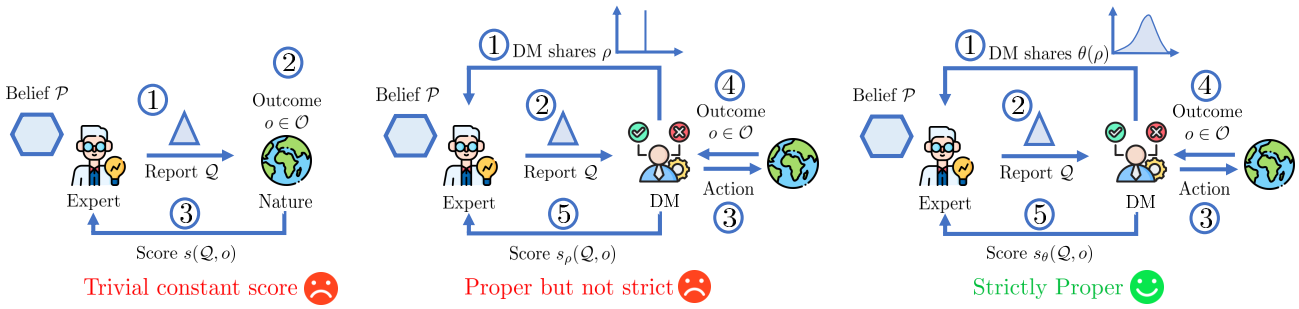


Figure 1: We consider scenarios where the expert holds an imprecise belief over the outcome $o \in \mathcal{O}$, represented as $\mathcal{P} \subseteq \Delta(\mathcal{O})$. The goal is to truthfully elicit this belief, i.e., the best report Q should be $\mathcal{P}$. The leftmost figure directly extends precise scoring rules to the imprecise case, ignoring the downstream DM. Truthful elicitation in the imprecise setting requires the DM to share their aggregation rule ρ with the expert (middle). To avoid DM's strategic manipulation by the forecaster, DM shares a distribution $\theta(\rho)$ over aggregation rules (right), resulting in a strictly proper scoring rule s_θ.

should be determined by the model's end user, also referred to as the decision-maker (DM). When direct interaction between the learner and the DM is not possible, Singh et al. [2024] propose an *imprecise learning* algorithm that trains a portfolio of predictors (forecasts) in advance, which are then provided to the DM. In contrast, for practical scenarios where the learner and DM can communicate, eliciting precise forecasts is straightforward using classical scoring rules. However, eliciting imprecise forecasts remains challenging due to the lack of suitable imprecise scoring rules. This gap motivates us to design appropriate imprecise scoring rules that are applicable beyond machine learning contexts.

The key challenge to designing an appropriate scoring rule arises from the forecaster's epistemic uncertainty. This challenge has led to several impossibility theorems for strictly proper imprecise scoring rules [Seidenfeld et al., 2012, Mayo-Wilson and Wheeler, 2015, Schoenfield, 2017]. However, these works focus solely on eliciting imprecise forecasts from the forecaster, overlooking the fact that probabilistic forecasts are typically used for downstream decision-making, making elicitation rarely the sole objective. Without input from the DM during elicitation, forecaster must rely solely on their imprecise belief, which contains inherent ambiguity. This often leads to indecision during elicitation—a key factor behind prior impossibility results. Recently, Fröhlich and Williamson [2024] explored imprecise scoring rules involving DMs, but their analysis focused only on min-max (pessimistic) decision-making and lacked formal discussion of the DM's role. More broadly, indecision can be resolved through subjective choices beyond the min-max rule. However, it cannot be resolved by forecasters alone without eliminating their epistemic uncertainty. We argue that the DM must actively assist forecasters in navigating indecision by communicating their subjective preferences.

Our contributions. To address this challenge, we propose a novel setup for scoring imprecise forecasts where we consider a DM as an additional agent, who actively guides the forecaster in resolving indecision during elicitation (see Figure 1 for different scenarios). Our contributions are summarized as follows:

- We show that prior impossibility results stem from the lack of communication between DM and the forecaster.
- We formalise DM-forecaster communication using aggregation rules from social choice theory [Arrow, 2012] and generalize tailored scoring rules [Johnstone et al., 2011] to accommodate these aggregations.
- We analyze the connection between axiomatic properties of aggregation rules from the social choice perspective and their impact on both truthful elicitation from the forecaster and the DM's decision-making process.
- By restricting to strategic communication, specifically by sharing only a distribution over aggregation rules, we propose a novel randomized tailored scoring rule that is strictly proper for imprecise forecasts.

The rest of the paper is organized as follows. Section 2 introduces proper scoring rules and imprecise probabilities. Section 3 then formalizes the notion of an imprecise forecaster and outlines decision-making for the forecaster and DM. Next, Section 4 explores imprecise scoring rules, first without communication and then with aggregation. Section 5 presents strictly proper scoring rules for imprecise forecasts, while Section 6 reviews prior work. Finally, Section 7 concludes with a discussion of future directions.

2 PRELIMINARIES

This section introduces proper scoring rules, imprecise probabilities (IP), and credal sets. We begin by establishing the notation. Let $(\mathcal{O}, \mathcal{F})$ be a measurable space where $\mathcal{O}$ is a finite, discrete, non-empty set of possible outcomes (or states of nature) and $\mathcal{F}$ is a corresponding sigma-algebra. Let $O : \mathcal{O} \to \mathbb{R}$ be a random variable associated with p.m.f. $p : \mathcal{O} \to [0, 1]$ on outcome $o \in \mathcal{O}$. The probability simplex $\Delta(\mathcal{O})$ denotes the set of all probability distributions on

$\mathcal{O}$. Our framework involves two agents: a forecaster and a decision-maker (DM), each with an associated utility function $u : \mathcal{X} \times \mathcal{O} \to \mathbb{R}$, where $\mathcal{X}$ represents the decision space relevant to the agent's utility. Since we often refer to specific outcomes $o \in \mathcal{O}$, we will use O and o interchangeability. Thus, for some $x \in \mathcal{X}$, the agent's expected utility $\mathbb{E}_{O\sim p}[u(x, O)]$ is expressed as $\mathbb{E}_{o\sim p}[u(x, o)]$. For a set $\mathcal{P}$, $co(\mathcal{P})$ corresponds to the convex hull and ext$(\mathcal{P})$ to its extreme points.

2.1 PRECISE SCORING RULES

Scoring rules incentivize a forecaster to truthfully report their probability assessments of an uncertain event [Winkler, 1967, Brier, 1950]. Specifically, a scoring rule $s : \Delta(\mathcal{O}) \times \mathcal{O} \to \mathbb{R}$ assigns a score of $s(q, o)$ to a forecaster with a forecast $q \in \Delta(\mathcal{O})$ when an outcome o happens.

Definition 2.1. A forecaster is precise if their true belief can be expressed as a probability distribution $p \in \Delta(\mathcal{O})$.

Since classical proper scoring rules focus on truthful reporting and evaluation of *precise* forecasts, we refer to them as precise scoring rules. To discourage a forecaster from making overly confident predictions, e.g., $q(o) = 0$. We introduce *regular* precise scoring rule, i.e. $s(q, o) \in \mathbb{R}$ for all $o \in \mathcal{O}$ and $s(q, o) = -\infty$ only if $q(o) = 0$.

Definition 2.2 (Expected utility of the forecaster)**.** Precise scoring rules implicitly assume that the forecaster is an expected utility-maximising agent. Therefore, for a forecaster with true belief p, the utility of reporting forecast q is

$$u_p(q) = \mathbb{E}_{o\sim p}[s(q, o)]. \tag{1}$$

We now define a subclass of regular precise scoring rules, known as *strictly* proper precise scoring rules that incentivize truthful reporting of the forecaster's belief.

Definition 2.3 (Strictly Proper Precise Scoring Rule)**.** A scoring rule $s : \Delta(\mathcal{O}) \times \mathcal{O} \to \mathbb{R} \cup \{-\infty\}$ is strictly proper if the forecaster's true belief $p \in \Delta(\mathcal{O})$ uniquely maximizes their expected utility, i.e., for all $p, q \in \Delta(\mathcal{O})$ s.t. $q \neq p$,

$$\mathbb{E}_{o\sim p}[s(p, o)] > \mathbb{E}_{o\sim p}[s(q, o)]. \tag{2}$$

Some examples of strictly proper precise scoring rules are, logarithmic scoring rule $s(q, o) = a_o + b \log(q(o))$ and quadratic scoring rule $s(q, o) = a_o + b(2q(o) - \mathbb{E}_{o\sim q}[q(o)])$ with $b \in \mathbb{R}_+$ and $a_o \in \mathbb{R}$ as arbitrary parameters. Proper precise scoring rules are closely related to convexity and can be characterized using convex functions as shown in McCarthy [1956], Savage [1971], Gneiting and Raftery [2007].

Theorem 2.4 (Gneiting and Raftery 2007)**.** *A regular scoring rule s is (strictly) proper if and only if*

$$s(q, o) = G(q) - \sum_{o\in\mathcal{O}} G'(q)dq(o) + G'(q)(o) \tag{3}$$

where $G : \Delta(\mathcal{O}) \to \mathbb{R}$ is a (strictly) convex function and $G'(q)$ is a subgradient of G at point q and $G'(q)(o)$ is the value of gradient at outcome o.

An implication of Theorem 2.4 is that with this characterisation of the scoring rule s, we can interpret G as the corresponding maximum expected score [Frongillo and Kash, 2014]. The derivation of G as the expected score is included in Appendix A.1 for completeness.

2.2 IP AND CREDAL SETS

Standard probability theory assigns a unique numerical value to each event, whereas *imprecise probabilities* (IP) allows a range of plausible values to represent uncertainty in the presence of limited or ambiguous information. One common approach to modelling such uncertainty is via *credal sets*. Given a subset $\mathcal{P} \subseteq \Delta(\mathcal{O})$ of the plausible probability distributions, a credal set is defined as a closed and convex combination of $\mathcal{P}$. The assumption of convexity and closedness allows for rational decision-making [Gajdos et al., 2004, Troffaes, 2007] and satisfies axioms such as coherence [de Finetti, 1974, Walley, 1991]. While $\mathcal{P}$ directly specifies the plausible beliefs about the state of nature, co$(\mathcal{P})$ denotes the uncertainty inferred by a rational agent [Walley, 1991, Augustin et al., 2014].

3 A JOINT DECISION FRAMEWORK FOR DM AND FORECASTER

In this work, we consider scenarios where an agent is tasked with selecting an input x from a finite space of inputs $\mathcal{X} := \{x_1, \ldots, x_n\}$. Agent's choice of input $x \in \mathcal{X}$ and outcome of uncertain event $o \in \mathcal{O}$ quantify the utility $u(x, o)$ obtained by the agent. In the case of a precise forecaster, $\mathcal{X} := \Delta(\mathcal{O})$ and Eq. (1) shows how the precise score $u(x, o) := s(p, o)$ acts as a utility for the forecaster, underlining the decision-making aspect within elicitation. From the DM's perspective, $\mathcal{X} := \mathcal{A}$ where $\mathcal{A} := \{a_1, \ldots, a_m\}$ denotes the finite space of actions which DM can choose from. Depending upon the outcome $o \in \mathcal{O}$, the DM obtains $u(x, o) := u(a, o)$ as the utility.

3.1 DECISION-MAKING WITH FORECASTS

There exists a crucial difference between decision-making with imprecise forecasts v.s. precise forecasts. In the case of precise forecasts, the agent (forecaster or DM) has a precise belief or report $p \in \Delta(\mathcal{O})$. Using p allows them to define a complete preference relation $\succeq_p$ over $\mathcal{X}$ based on several well-established rationality frameworks [Von Neumann and Morgenstern, 1947, Savage, 1972]. Thereby, allowing the agent to select the corresponding best input x^*. This x^* represents the best forecast to report in the case of a precise

forecaster and the best action to take in the case of DM. However, in scenarios where the belief (or obtained report) for an agent is a set of presice beliefs $\mathcal{P} \subseteq \Delta(\mathcal{O})$, the preference relation ($\succeq_{\mathcal{P}}$) obtained on $\mathcal{X}$ using $\mathcal{P}$ is incomplete. In this case, a natural way to define $\succeq_{\mathcal{P}}$ is based on the idea of dominance.

Definition 3.1. Consider $\mathcal{P} \subseteq \Delta(\mathcal{O})$, then the corresponding preference relation $\succeq_{\mathcal{P}}$ over $\mathcal{X}$ for a VNM rational [Von Neumann and Morgenstern, 1947] agent can be defined as follows: for all $x, x' \in \mathcal{X}$,

$$x \succeq_{\mathcal{P}} x' \quad \text{iff} \quad \mathbb{E}_p[u(x,o)] \geq \mathbb{E}_p[u(x',o)] \quad \forall p \in \mathcal{P}.$$

Unless $\mathcal{P}$ is implicitly a precise forecast of type $\{p \in \Delta(\mathcal{O})\}$, the preference relation $\succeq_{\mathcal{P}}$ is a partial order over $\mathcal{X}$. The partial order $\succeq_{\mathcal{P}}$ can be incomplete, since there can be a pair of inputs $x, x' \in \mathcal{X}$ such that $x' \not\succeq_{\mathcal{P}} x$ and $x \not\succeq_{\mathcal{P}} x'$. In other words, x and x' are incomparable. This can result in indecision for the agent. This means that both the forecaster and the DM face indecision when they rely on $\mathcal{P}$ for their respective tasks (elicitation or decision-making).

3.2 IMPRECISE FORECASTER

Our work focuses on analyzing scoring rules in scenarios where the forecaster may be *imprecise*. Specifically, we formalise the notion of an imprecise forecaster and their truthfulness below.

Definition 3.2. A forecaster is imprecise if their belief can be expressed as a set of distributions $\mathcal{P} \subseteq \Delta(\mathcal{O})$. A report $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ is called an imprecise forecast, which implicitly includes precise forecasts $\mathcal{Q} = \{q\}$ for some $q \in \Delta(\mathcal{O})$.

Definition 3.2 generalizes the precise setting as it allows the forecaster to express their (partial) ignorance by reporting both aleatoric uncertainties (as elements in the set) and epistemic uncertainties (as the set itself) [Hüllermeier and Waegeman, 2021]. This subsumes both scenarios where the forecaster's belief is truly imprecise, e.g., the probability that it will rain tomorrow is $[0.6, 0.8]$, and where their belief is calibrated with respect to multiple sources of potentially conflicting information, e.g., the estimated probability based on data from multiple weather stations. Moreover, this can also be interpreted as a "collective" report obtained from multiple (potentially conflicting) precise forecasters. Imprecise probability scoring rules can be defined analogously to precise scoring rules as follows.

Definition 3.3. (Imprecise Probability Scoring Rule) An imprecise probability (IP) scoring rule $s : 2^{\Delta(\mathcal{O})} \times \mathcal{O} \to \mathbb{R} \cup \{-\infty\}$ assigns a score of $s(\mathcal{Q}, o)$ to a report $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ when the outcome $o \in \mathcal{O}$ is realized.

Analogous to precise setting, an IP scoring rule is *regular* if $s(\mathcal{Q}, o) \in \mathbb{R}$ for all $o \in \mathcal{O}$, except if $q(o) = 0$ for all $q \in \mathcal{Q}$, then $s(\mathcal{Q}, o) = -\infty$. To define regularity analogous to the precise setting we consider for all $q \in \mathcal{Q}$, since otherwise reporting a vacuous set $\Delta(\mathcal{O})$ or other imprecise sets will have $-\infty$ as an incentive, thereby discouraging the forecaster from reporting their epistemic uncertainty. The score $s(\mathcal{Q}, o)$ obtained by the forecaster induces a corresponding set of utilities $\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})$ for the forecaster with an imprecise belief $\mathcal{P}$, representing the expected utility of the imprecise score with respect to every distribution within their belief $\mathcal{P}$. We define this utility set as follows:

$$\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q}) = \{\mathbb{E}_p[s(\mathcal{Q}, o)]\}_{p \in \mathcal{P}}$$

From the forecaster's perspective, this collection of expected utility functions $\boldsymbol{V}^{\mathcal{P}} : 2^{\Delta(\mathcal{O})} \to \mathbb{R}^{|\mathcal{P}|}$, for each report $\mathcal{Q}$ result in a range of plausible expected utility, i.e.,

$$\mathrm{im}(\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})) = \left[\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)], \sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] \right]$$

where im is the image or the range of the forecaster's minimum and maximum expected score for forecast $\mathcal{P}$ when its extreme points exist; see Appendix A.2 for further details. While the equivalence of two precise distributions p and q is natural, i.e., $p = q$ or not. The equivalence of two imprecise beliefs is not obvious as they are sets of distributions. We now define the equivalence of two beliefs $\mathcal{P}, \mathcal{P}'$ in the context of elicitation as follows.

Definition 3.4. (Equivalence of Imprecise Beliefs) Two beliefs $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{O})$ are considered equivalent, denoted as $\mathcal{P} \simeq \mathcal{P}'$, if for all IP scoring rules s and forecasts $\mathcal{Q} \subseteq \Delta(\mathcal{O})$, we have $\mathrm{im}(\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})) = \mathrm{im}(\boldsymbol{V}^{\mathcal{P}'}(\mathcal{Q}))$.

Intuitively, two imprecise forecasts are equivalent if they yield the same range of plausible expected utilities for any scoring rule s and reported forecast $\mathcal{Q}$—that is, they induce identical decision-making. We now show that Definition 3.4 reduces to the classic notion of equivalence between probability distributions when applied to precise forecasts.

Proposition 3.5. *For all $p, q \in \Delta(\mathcal{O})$, $\{p\} \simeq \{q\}$ iff $p = q$.*

With Proposition 3.5, we establish that Definition 3.4 generalises from the notion of equivalence of precise forecasts, i.e. distributions to imprecise forecasts. We can also characterize the equivalence of two imprecise forecasts as the equivalence of their corresponding credal sets.

Proposition 3.6. *For imprecise beliefs $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{O})$ with non-empty extreme points, $\mathcal{P} \simeq \mathcal{P}'$ iff $co(\mathcal{P}) = co(\mathcal{P}')$.*

It has previously been shown that two sets of distributions must be credal sets to induce the same rational decision-making behaviour [Troffaes, 2007, Huntley et al., 2014, Troffaes and de Cooman, 2014]. Definition 3.4 defines the

equivalence of two imprecise beliefs w.r.t elicitation and Proposition 3.6 establishes its equivalence to rational decision making. This allows us to consider elicitation as a decision-making task for the forecaster. As a consequence of Proposition 3.6, even though a forecaster believes in a set of probability distributions $\mathcal{P}$. We restrict our focus to evaluating a credal set of forecasts $co(\mathcal{P})$. Therefore, from now on, we will assume that $\mathcal{P}$ is a convex set.

Definition 3.7. (Truthfulness of Imprecise Forecaster) Let $\mathcal{P} \subseteq \Delta(\mathcal{O})$ be the true belief of an imprecise forecaster. A report $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ is truthful if $\mathcal{Q} \simeq \mathcal{P}$.

Definition 3.7 generalizes the concept of truthfulness in the precise setting. An imprecise forecaster who reports their true belief is considered truthful. For instance, if the forecaster believes the probability of rain tomorrow lies within the interval $[0.6, 0.8]$, then they must report their actual epistemic uncertainty by reporting the interval $[0.6, 0.8]$.

4 PROPER IP SCORING RULES

In this section, we introduce proper imprecise scoring rules, i.e., scores that incentivize the truthful reporting of an imprecise forecaster according to Definition 3.7. We start by focusing only on the elicitation of the imprecise forecaster without any communication from the DM. The following definition clarifies what it means for an imprecise scoring rule to be (strictly) proper, which naturally generalises Definition 3.1 from the forecaster's perspective.

Definition 4.1. An imprecise scoring rule is (strictly) proper if for all credal sets $\mathcal{P}, \mathcal{Q} \subseteq \Delta(\mathcal{O})$, the forecaster with an imprecise belief $\mathcal{P} \neq \mathcal{Q}$ (strictly) prefers $\mathcal{P}$ over $\mathcal{Q}$, i.e., $\mathcal{P} \succeq_{\mathcal{P}} \mathcal{Q}$. The preference relation $\succeq_{\mathcal{P}}$ is described by the (strict) dominance of $\boldsymbol{V}^{\mathcal{P}}(\mathcal{P})$ over $\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})$, i.e.,

$$\mathbb{E}_p[s(\mathcal{P}, o)] \geq \mathbb{E}_p[s(\mathcal{Q}, o)] \quad \text{for all } p \in \mathcal{P},$$

for strict dominance, at least one $\geq$ has to be strictly greater.

We define strict properness of an imprecise scoring rule in Definition 4.1 using dominance since it preserves the main idea behind strictly proper scoring rules in the precise setting, i.e. to incentivise the forecaster to be truthful. A strictly proper IP scoring rule incentivises the imprecise forecaster to be truthful according to Definition 3.7.

Theorem 4.2. *There does not exist a strictly proper imprecise scoring rule s. In addition, for a scoring rule s to be proper it must be constant across all forecasts.*

Similar impossibility results for imprecise forecasts have previously been reported in Seidenfeld et al. [2012], Mayo-Wilson and Wheeler [2015], Schoenfield [2017]. The implication of Theorem 4.2 is that under the current setup of an imprecise forecaster, any imprecise scoring rule satisfying properness in Definition 4.1 has a constant score across all forecasts. We observe in Section 3.1 that agents face possible indecision while making decision with the credal set $\mathcal{P}$. As a result, we observe in Theorem 4.2 that it is not possible to design a scoring rule that incentivises the imprecise forecaster to report their imprecise belief $\mathcal{P}$ honestly. Unlike in the precise setting, where the forecaster had a complete preference relation on plausible reports (see Definition 2.3), the epistemic uncertainty of the imprecise forecaster only allows for an incomplete preference relation $\succeq_{\mathcal{P}}$ over plausible reports. Without further information, the imprecise forecaster cannot complete this incomplete preference relation.

4.1 AGGREGATION FUNCTIONS

To resolve indecision arising from epistemic uncertainty in the credal set $\mathcal{P}$, the DM exercises a subjective choice through aggregation function ρ to make $\succeq_{\mathcal{P}}$ complete. The DM communicates the choice of ρ to the forecaster *prior* to elicitation, and the elicited credal set then informs downstream decisions for the DM. The resulting utility can be shared with the forecaster as an incentive.

Definition 4.3 (Aggregation Function)**.** For a credal set $\mathcal{Q}$ an aggregation function $\rho : (\mathbb{R}^{\mathcal{X}})^{|\mathcal{Q}|} \to \mathbb{R}^{\mathcal{X}}$ combines multiple utilities via a positive linear combination, i.e., for any $x \in \mathcal{X}$:

$$\rho[\{\mathbb{E}_q[u(x, o)]\}_{q \in \mathcal{Q}}] = \int_{q \in \mathcal{Q}} \boldsymbol{w}(q) \mathbb{E}_q[u(x, o)] dq$$

where $\boldsymbol{w}(q) \in \mathbb{R}^{|\mathcal{Q}|}_{\geq 0}$ for all $q \in \mathcal{Q}$ depends on the expected utilities $\{\mathbb{E}_q[u(x, o)]\}_{q \in \mathcal{Q}}$.

We focus on linear aggregations because, according to Harsanyi [1955], this class of aggregation rules uniquely satisfies both VNM axioms [Von Neumann and Morgenstern, 1947] and Bayes Optimality [Brown, 1981]. Many popular aggregation functions such as utilitarian and egalitarian rules can be expressed with linear aggregations as they characterise relative utilitarianism [Dhillon and Mertens, 1999].

For the utilitarian and egalitarian rules, the decision-making process from an agent's perspective is picking an $x \in \mathcal{X}$. Either for the DM an $a \in \mathcal{A}$ or for the forecaster a $\mathcal{Q} \subseteq \Delta(\mathcal{O})$. Illustrating this from the DM's perspective, the utilitarian rule corresponds to the linear combination $\rho[\{\mathbb{E}_q[u(a, o)]\}_{q \in \mathcal{Q}}] = 1/|\mathcal{Q}| \sum_{q \in \mathcal{Q}} \mathbb{E}_q[u(a, o)]$, whereas the egalitarian rule corresponds to $\rho[\{\mathbb{E}_q[u(a, o)]\}_{q \in \mathcal{Q}}] = \min_{q \in \mathcal{Q}} \mathbb{E}_q[u(a, o)]$. Here, the weights $\boldsymbol{w}$ can be interpreted as $\boldsymbol{w}(q) = 1/|\mathcal{Q}|$ and a one-hot vector, respectively. A VNM-rational DM (see Eq. (5) for forecaster) uses ρ to obtain the best action $a^*_{\mathcal{Q},\rho}$:

$$a^*_{\mathcal{Q},\rho} = \arg\max_{a \in \mathcal{A}} \rho[\{\mathbb{E}_q[u(a, o)]\}_{q \in \mathcal{Q}}]. \tag{4}$$

Given the incomplete preference relation from a credal set $\mathcal{Q}$, i.e., $\succeq_{\mathcal{Q}} := \{\succeq_q\}_{q\in\mathcal{Q}}$, the aggregation rule ρ allows us to define the corresponding complete preference relation $\succeq_{\rho[\mathcal{Q}]}$, representing the aggregated utility from Equation (4). By abuse of notation, $\succeq_{\rho[\mathcal{Q}]}$ represents the aggregation of utilities rather than the credal set.

Axiomatisation of ρ: When interpreting imprecise forecasts as a "collective" report of precise forecasters, a social choice perspective naturally emerges for the downstream DM. Although non-intuitive, this perspective applies even to a single-agent imprecise forecaster. Following Arrow [1950], we outline three desirable properties of any aggregation rule ρ: Pareto Efficiency (PE), Independence from Irrelevant Alternatives (IIA), and Non-Dictatorship (ND).

Definition 4.4 (Pareto Efficiency)**.** An aggregation rule ρ is Pareto Efficient iff for all $x, x' \in \mathcal{X}$,

$$x \succeq_{\mathcal{Q}} x' \implies x \succeq_{\rho[\mathcal{Q}]} x'.$$

From the DM's perspective, $\mathcal{X} = \mathcal{A}$, and as a result, a Pareto efficient ρ will respect the inherent partial order $\succeq_{\mathcal{Q}}$ over actions which DM could infer from the reported credal set $\mathcal{Q}$. Therefore, the DM can be assured that application of ρ only resolves indecision and similarly for the forecaster when choosing the best report. Additionally, from the forecaster's perspective, a non-PE ρ can distort recommendations of their forecast $\mathcal{Q}$ of an action a over a'. The aggregation rule that violates PE may result in the payment/score that misaligns with the forecaster's report.

Definition 4.5 (IIA)**.** An aggregation function ρ is considered IIA if preferences between $x, y \in \mathcal{X}$, i.e., $x \succeq_{\rho[\mathcal{Q}]} y$ or $y \succeq_{\rho[\mathcal{Q}]} x$ is independent of whether any other z is in $\mathcal{X}$.

Although cryptic, IIA is desirable to the DM. From the DM's perspective, $\mathcal{X} = \mathcal{A}$. Imagine a scenario where there exists a $z \in \mathcal{A}$ such that both $x, y \in \mathcal{A}$ dominate z w.r.t. the partial order $\succeq_{\mathcal{Q}}$, implying that z is irrelevant to the DM under forecasts $\mathcal{Q}$. However, if ρ violates IIA, the post-aggregation preference $\succeq_{\rho[\mathcal{Q}]}$ between x and y can be influenced by the presence or absence of z. This makes the DM vulnerable to strategic manipulation regarding the best action to take by adding or removing z, which in turn creates uncertainty for the forecaster about their own incentives.

Definition 4.6 (Non-Dictatorship)**.** An aggregation rule ρ is said to be non-dictatorial if for a profile of preferences $\succeq_{\mathcal{Q}} := \{\succeq_q\}_{q\in\mathcal{Q}}$ there does not exist $q \in \mathcal{Q}$ (dictator) such that for all $x, y \in \mathcal{X}$, $x \succeq_q y$ implies $x \succeq_{\rho[\mathcal{Q}]} y$.

From the downstream decision-making perspective for a DM, non-dictatorship is optional. However, when DM wants to communicate the aggregation rule ρ to the forecaster and wishes to truthfully elicit their true belief, non-dictatorship becomes crucial. Given a dictatorial ρ, the forecaster can manipulate the DM by strategically reporting the dictator q. We discuss this more formally in Appendix C.1.

4.2 PROPER IP SCORES WITH AGGREGATION

The DM communicates the aggregation function ρ to the forecaster and incentivises them using an IP scoring rule. This communication helps resolve the forecaster's epistemic uncertainty, parameterizing the IP scoring rule as $s_\rho : 2^{\Delta(\mathcal{O})} \times \mathcal{O} \to \mathbb{R}$. The forecaster reports $\mathcal{Q} \in 2^{\Delta(\mathcal{O})}$ and receives a score of $s_\rho(\mathcal{Q}, o)$ when outcome $o \in \mathcal{O}$ occurs. Unlike prior IP scoring rules, the forecaster can now use ρ to resolve indecision and complete the preference relation over $2^{\Delta(\mathcal{O})}$. This is evident from the expected utility of the forecaster with belief $\mathcal{P}$ when reporting $\mathcal{Q} \subseteq \Delta(\mathcal{O})$:

$$V_\rho^{\mathcal{P}}(\mathcal{Q}) := \rho[\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})] = \rho[\{\mathbb{E}_p[s_\rho(\mathcal{Q}, o)]\}_{p\in\mathcal{P}}]. \quad (5)$$

Since an imprecise decision scoring rule s_ρ is simply a parameterised IP scoring rule, its regularity is defined exactly as in Section 4. We extend (strict) properness for IP scoring rules from Definition 2.3 to aggregation as follows.

Definition 4.7. A regular IP scoring rule s_ρ for an aggregation function ρ is proper if, for all $\mathcal{P} \subseteq \Delta(\mathcal{O})$ and all $\mathcal{Q} \neq \mathcal{P}$, $V_\rho^{\mathcal{P}}(\mathcal{P}) \geq V_\rho^{\mathcal{P}}(\mathcal{Q})$. The IP scoring rule s_ρ is strictly proper if and only if at least one of the inequalities is strict.

Notably, strictness in Definition 4.7 adheres to the notion of truthfulness defined in Definition 3.7. Since DM needs to evaluate the forecaster, we employ the class of scoring rules that accommodate a DM in evaluating a forecast, called tailored scoring rules [Dawid, 2007, Richmond et al., 2008, Johnstone et al., 2011], following the ideas of business sharing proposed in Savage [1971]. We now define them in the context of aggregation functions for imprecise forecasts.

Definition 4.8 (Tailored Scoring Rules)**.** An IP scoring rule s is tailored for a DM with utility function u and aggregation function ρ, if for any $k, c \in \mathbb{R}_{\geq 0}$, the score is defined as

$$s_\rho(\mathcal{Q}, o) = ku(a^*_{\mathcal{Q},\rho}, o) + c.$$

In Definition 4.8, k can be referred to as the business share obtained by the forecaster in the utility of the DM and c is the fixed fee of the forecaster. Next, we show that the class of tailored scoring rules is proper for any choice of ρ,

Proposition 4.9. *A tailored scoring rule s_ρ is proper with respect to Definition 4.7 for any aggregation rule ρ.*

While necessary, the properness of scoring rules is easy to satisfy (see Theorem 4.2). For example, a constant scoring rule is always proper. We therefore characterise strict properness of s_ρ for imprecise forecasts.

Lemma 4.10. *Let s_ρ be a tailored scoring rule. Then, the following holds:*

1. *s_ρ is strictly proper for **precise distributions** if and only if $a^*_q := \arg\max_{a\in\mathcal{A}} \mathbb{E}_q[u(a, o)]$ is a unique maximiser for all $q \in \Delta(\mathcal{O})$.*

2. *s_ρ is not strictly proper, i.e., does not satisfy Definition 4.7, for any Pareto efficient ρ.*

Lemma 4.10 ensures the existence of non-constant proper IP scoring rules. Beyond this positive result, we observe that Pareto efficiency leads to the impossibility of truthful elicitation under Definition 3.7. Although s_ρ in Lemma 4.10 is not strictly proper for imprecise forecasts, it remains practical to implement while being proper for all forecasts and strictly proper for precise ones. We speculate that the properties of s_ρ are optimal for deterministic scoring rules, given the prior impossibility of any real-valued strictly proper IP scoring rules [Seidenfeld et al., 2012]. To explore this further, we investigate whether allowing randomisation in the choice of aggregation rule can enable truthful elicitation.

5 STRICTLY PROPER IP SCORES

With the randomized choice of aggregation, the DM can pick an aggregation rule randomly post-elicitation to evaluate the reported forecast. The forecaster then becomes unaware of the aggregation function which can lead the forecaster back to indecision. To resolve the forecaster's indecision, the DM shares a distribution $\theta \in \Delta(\boldsymbol{\rho})$ where $\boldsymbol{\rho}$ is the class of aggregation functions the DM will pick from, thereby enabling the forecaster to resolve their indecision as follows:

$$V_\theta^{\mathcal{P}}(\mathcal{Q}) := \mathbb{E}_{\rho\sim\theta}[V_\rho^{\mathcal{P}}(\mathcal{Q})]. \tag{6}$$

This allows the tailored scoring rule s_ρ to be randomized with respect to the random variable ρ. Analogous to Definition 4.8 for tailored scoring rules, we now define randomized tailored scoring rule s_θ.

Definition 5.1. A regular IP scoring rule s_θ is randomized tailored for a DM with a class of aggregation functions $\boldsymbol{\rho}$ and a distribution $\theta \in \Delta(\boldsymbol{\rho})$, if for any $k_\rho, c_\rho \in \mathbb{R}_{\geq 0}$ and an arbitrary function $\Pi : 2^{\Delta(\mathcal{O})} \to \mathbb{R}$, the score is defined as

$$s_\theta(\mathcal{Q}, o)(\rho) = \begin{cases} k_\rho u(a^*_{\rho,\mathcal{Q}}, o) + c_\rho & \text{if } \theta(\rho) > 0 \\ \Pi_o(\mathcal{Q}) & \text{if } \theta(\rho) = 0 \end{cases}.$$

Given that we have now extended the tailored scoring rule to random variables, in a similar spirit to Definition 4.7 on properness of IP scoring rules with aggregation, we define properness of randomized tailored scoring rules as follows.

Definition 5.2. A randomized tailored scoring rule s_θ for a distribution $\theta \in \Delta(\boldsymbol{\rho})$ and a class of aggregation rules $\boldsymbol{\rho}$, is considered proper if, for all $\mathcal{P}, \mathcal{Q} \subseteq \Delta(\mathcal{O})$ and $\mathcal{Q} \neq \mathcal{P}$,

$$V_\theta^{\mathcal{P}}(\mathcal{P}) \geq V_\theta^{\mathcal{P}}(\mathcal{Q}). \tag{7}$$

s_θ is strictly proper if the inequality in Equation (7) is strict.

Again the strictness in Definition 5.2 adheres to the notion of truthfulness defined in Definition 3.7. We will establish this connection later in this section. We can observe from Equation 6 that randomized tailored scoring rules are proper for any choice of $\theta \in \Delta(\boldsymbol{\rho})$ as a direct consequence of Proposition 4.9. Before we discuss how to build strictly proper IP scoring rules, we need to identify if there exists a unique representation of the credal set in the action space which will let the DM identify the credal set.

Lemma 5.3. *For any reported credal set $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ and a DM using a utility function u such that $a_q^* := \arg\max_{a\in\mathcal{A}} \mathbb{E}_q[u(a, o)]$ is unique for all $q \in \Delta(\mathcal{O})$, the set of actions $\mathcal{A}_{\mathcal{Q}}^{ext} := \left\{a_q^*\right\}_{q\in ext(\mathcal{Q})}$ acts as a unique representation of a credal set $\mathcal{Q}$ in action space $\mathcal{A}$.*

The implication of unique representation $\mathcal{A}_{\mathcal{Q}}^{\text{ext}}$ in the action space for any credal set $\mathcal{Q}$ is that the DM is able to identify the credal set from the set of actions $\mathcal{A}_{\mathcal{Q}}^{\text{ext}}$. In a naive analogy, all actions in $\mathcal{A}_{\mathcal{Q}}^{\text{ext}}$ together act as a fingerprint of credal set $\mathcal{Q}$ which can be uniquely incentivised by the DM to elicit $\mathcal{Q}$. We now introduce a common class of linear aggregations to operationalise scoring rules based on Lemma 5.3.

Fixed Linear Aggregations is another common class of aggregation rules which aggregates the expected utilities of a credal set $\mathcal{Q}$ for any input $x \in \mathcal{X}$, i.e., $\{\mathbb{E}_q[u(x, o)]\}_{q\in\mathcal{Q}}$, into a convex combination of utilities with mixing coefficient $\boldsymbol{\lambda} \in \Delta^{|\mathcal{Q}|}$ as

$$\begin{aligned}\rho_{\boldsymbol{\lambda}}[\{\mathbb{E}_q[u(x, o)]\}_{q\in\mathcal{Q}}] &:= \int_{q\in\mathcal{Q}} \boldsymbol{\lambda}(q)\mathbb{E}_q[u(x, o)]dq \\ &= \mathbb{E}_{\int \boldsymbol{\lambda}(q)qdq}[u(x, o)].\end{aligned}$$

Although the class of fixed linear aggregations is Pareto-efficient and non-dictatorial in classic social choice theory, in our setup, fixed linear aggregations are dictatorships as they directly aggregate the epistemic uncertainty. Due to Proposition 3.6, a forecaster can report $\mathcal{Q}$ or $\text{co}(\mathcal{Q})$. We illustrate this with an example, for any report $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ and any choice of fixed linear aggregation $\boldsymbol{\lambda}$, we obtain $Q := \boldsymbol{\lambda}^\top \mathcal{Q}$. Even though Q may not be in $\mathcal{Q}$, it is guaranteed that $Q \in \text{co}(\mathcal{Q})$, and therefore Q acts as a dictator. This means that although the DM uses the full credal set in the sense of all extreme points to perform decision-making, their preference over actions can be fully represented by a precise belief $P \in co(\mathcal{P})$. From Section 4.1, non-dictatorship was only desirable due to the strategic manipulation by the forecaster. In the scenario where forecasters are unaware of the exact aggregation rule, using a random dictatorial $\rho_{\boldsymbol{\lambda}}$ allows the DM to keep PE and IIA. To this end, we show the strict properness of these randomized dictatorships. Since strict properness for imprecise forecasters implicitly requires strictness for precise forecasts, which means that the s_θ must satisfy Lemma 4.10 for every $\boldsymbol{\lambda}$.

Theorem 5.4. *Assuming s_θ to be strictly proper for precise distributions and $\boldsymbol{\rho}$ as fixed linear aggregations, s_θ is strictly proper for imprecise forecasts, i.e. s_θ is a strictly proper IP scoring rule if θ has full support over $\boldsymbol{\rho}$.*

Theorem 5.4 allows us to build strictly proper IP scoring rules which can be characterized as follows. A randomized tailored scoring rule s_θ made using the class of fixed linear aggregation rules is characterized as

$$s_\theta(\mathcal{Q}, o)(\boldsymbol{\lambda}) = \begin{cases} k_{\boldsymbol{\lambda}} u(a^*_{\rho_\lambda, \mathcal{Q}}, o) + c_{\boldsymbol{\lambda}} & \text{if } \theta(\boldsymbol{\lambda}) > 0 \\ \Pi_o(\mathcal{Q}) & \text{if } \theta(\boldsymbol{\lambda}) = 0 \end{cases},$$

where $\boldsymbol{\lambda} \in \Delta^{|\text{ext}(\mathcal{Q})|}$ is considered strictly proper if $\text{supp}(\theta) = [0, 1]$ where $\Pi : 2^{\Delta(\mathcal{O})} \times \mathcal{O} \to \mathbb{R}$ is an arbitrary regular scoring function. To verify the strict properness of our score, we conduct a simulation (see Appendix E).

In recent years, several frameworks have been proposed for learning that challenge the implicit assumptions made in standard ML pipeline about loss functions [Gopalan et al., 2021]) or preferences [Singh et al., 2024] of the users being known to the model trainer. They focus on training models that perform well for a class of losses or aggregation rules. Within our setup, these frameworks translate to the DM abstaining from sharing the exact aggregation rule with the forecaster. However, they are not exact implementations of the score we propose. Applying the proposed score to ML applications is one of the future research avenues.

6 RELATED WORK

The work of Fröhlich and Williamson [2024] is most closely related to ours. They also explore the generalization of proper scoring rules to imprecise forecasts, with a specific emphasis on calibration [Dawid, 1982]. While their focus is on imprecisions arising from data models, we address more general issues related to the elicitation of imprecise forecasts. Their findings demonstrate that, unlike in precise settings where proper scoring and calibration objectives align, these goals can diverge when dealing with imprecise forecasts—a result that parallels our own. However, their reliance on the min-max aggregation within their scoring framework limits their analysis to pessimistic decision-making, resulting in a scoring rule that only satisfies properness.

Impossibility results show that no continuous scoring rule over credal sets can satisfy strict incentive compatibility, calibration, and non-domination simultaneously [Seidenfeld et al., 2012, Mayo-Wilson and Wheeler, 2015, Schoenfield, 2017]. Seidenfeld et al. [2012] proved that such rules must either weaken incentive compatibility or permit domination by precise forecasts. Mayo-Wilson and Wheeler [2015] highlighted that these trade-offs can inadvertently reward false precision, while Schoenfield [2017] showed that any continuous rule is either constant or fails to calibrate in natural decision contexts. While our approach partly mitigates these issues, these impossibility results still constrain deterministic methods. Some view the lack of imprecise scoring rules analogous to precise ones as a fundamental trait of imprecision [Konek, 2015]. Building on this, Konek [2019] proposes a family of IP scoring rules based on the Hurwicz criterion, extended by Konek [2023] to formalize precision–robustness trade-offs axiomatically. Since the Hurwicz criterion yields Pareto-efficient aggregation, our results in Section 4 directly apply to their framework, offering a social choice lens on these trade-offs.

Finally, our work is uniquely positioned at the intersection of proper scoring rules, forecast elicitation, and machine learning, providing novel perspectives on decision-making under uncertainty. Credal sets have become a mainstream approach for representing modelers' imprecision with applications in prediction [Singh et al., 2024, Caprio et al., 2024], uncertainty quantification [Sale et al., 2023, Wang et al., 2024], optimal transport [Caprio, 2024], statistical hypothesis testing [Chau et al., 2025a], and statistical distances [Chau et al., 2025b], among others. To this end, our results concerning strictly proper scoring rules for credal sets are directly relevant to the challenges of learning and decision-making with credal sets, providing insights into fundamental problems and future research directions.

7 DISCUSSION

Our investigation of strictly proper IP scoring rules reveals that, unlike in the classical precise setting, forecasting under imprecision demands careful attention to the decision-making aspect within forecast elicitation. In traditional frameworks with strictly proper scoring rules, forecasters are simply expected utility maximizers, making the reporting decision straightforward. However, when forecasts are imprecise—represented as sets or intervals—forecasters cannot internally aggregate their epistemic uncertainty. Instead, they require an external aggregation rule to reconcile their credal set-induced preferences into a single forecast.

This need for external decision guidance naturally connects to social choice theory. In our framework, the DM provides a collective aggregation rule that guides forecasters in resolving their uncertainty. This approach not only preserves incentive compatibility in the imprecise setting but also highlights the importance of designing scoring rules that balance accuracy and robustness. By explicitly integrating a social choice–inspired aggregation function into the elicitation process, our work offers new perspectives on collective decision-making, where imprecise forecasts can be viewed as forecasts of the "collective." This highlights promising directions for future research on imprecise scoring rules.

Acknowledgement We thank the anonymous reviewers for their insightful feedback.

References

Pierre Pinson and Robin Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012.

Pierre Pinson. Wind energy: Forecasting challenges for its operational management. 2013.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1190–1205. PMLR, 2022.

Hiroki Yanagisawa. Proper scoring rules for survival analysis. In *International Conference on Machine Learning*, pages 39165–39182. PMLR, 2023.

Anurag Singh, Mahalakshmi Sabanayagam, Krikamol Muandet, and Debarghya Ghoshdastidar. Robust feature inference: A test-time defense strategy using spectral projections. *arXiv preprint arXiv:2307.11672*, 2023.

Chenze Shao, Fandong Meng, Yijin Liu, and Jie Zhou. Language generation with strictly proper scoring rules. *arXiv preprint arXiv:2405.18906*, 2024.

Yifan Wu and Jason Hartline. Elicitationgpt: Text elicitation mechanisms via language models. *arXiv preprint arXiv:2406.09363*, 2024.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Allan H. Murphy and Robert L. Winkler. A new vector partition of the probability score. *Journal of Applied Meteorology*, 27:1200–1207, 1988.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Rafael Frongillo and Ian Kash. General truthfulness characterizations via convex analysis. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings 10*, pages 354–370. Springer, 2014.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831. URL https://doi.org/10.1146/annurev-statistics-062713-085831. _eprint: https://doi.org/10.1146/annurev-statistics-062713-085831.

Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 3rd edition, 2011.

Jason Konek. Epistemic conservativity and imprecise credence. *Preprint*, 2015. URL https://philpapers.org/rec/KONECA.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain Generalization via Invariant Feature Representation, January 2013. URL http://arxiv.org/abs/1301.2115. arXiv:1301.2115 [cs, stat].

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, apr 2023.

Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. *arXiv preprint arXiv:2404.04669*, 2024.

Teddy Seidenfeld, Mark J Schervish, and Joseph B Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53(8):1248–1261, 2012.

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

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

Christian Fröhlich and Robert C Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024.

Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

David J Johnstone, Victor Richmond R Jose, and Robert L Winkler. Tailored scoring rules for probabilities. *Decision Analysis*, 8(4):256–268, 2011.

Robert L Winkler. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, 62(320):1105–1120, 1967.

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9): 654–655, 1956.

Thibault Gajdos, Jean-Marc Tallon, and Jean-Christophe Vergnaud. Decision making with imprecise probabilistic information. *Journal of Mathematical Economics*, 40(6): 647–681, 2004.

Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45(1):17–29, 2007.

Bruno de Finetti. *Theory of Probability*. John Wiley & Sons, 1974.

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

Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to Imprecise Probabilities*. John Wiley & Sons, Chichester, 2014.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1947.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3): 457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://arxiv.org/abs/1910.09457. arXiv:1910.09457 [cs, stat].

Nathan Huntley, Robert Hable, and Matthias CM Troffaes. Decision making. *Introduction to imprecise probabilities*, pages 190–206, 2014.

Matthias C. M. Troffaes and Gert de Cooman. Lower previsions. In *Introduction to Imprecise Probabilities*, pages 159–181. John Wiley & Sons, Chichester, 2014.

John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955. URL https://www.jstor.org/stable/1827128.

Lawrence D Brown. A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics*, pages 1289–1300, 1981. URL https://www.jstor.org/stable/2240418.

Amrita Dhillon and Jean-François Mertens. Relative utilitarianism. *Econometrica*, 67(3):471–498, 1999.

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

A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

Jose Victor Richmond, Robert F Nau, and Robert L Winkler. Scoring rules, generalized entropy, and utility maximization. *Operations research*, (5):1146–1157, 2008.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Jason Konek. Ip scoring rules: Foundations and applications. In *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, pages 256–264. PMLR, 2019. URL https://proceedings.mlr.press/v103/konek19a.html.

Jason Konek. Evaluating imprecise forecasts. In *Proceedings of the International Symposium on Imprecise Probability: Theories and Applications*, volume 215, pages 270–279. PMLR, 2023. URL https://proceedings.mlr.press/v215/konek23a.html.

Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal Learning Theory, February 2024. URL http://arxiv.org/abs/2402.00957. arXiv:2402.00957 [cs, stat].

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].

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation on out-of-distribution detection. *arXiv preprint arXiv:2405.15047*, 2024.

Michele Caprio. Optimal transport for ϵ-contaminated credal sets. *arXiv preprint arXiv:2410.03267*, 2024.

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, 2025a.

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

Errett Bishop and Karel De Leeuw. The representations of linear functionals by measures on sets of extreme points. In *Annales de l'institut Fourier*, volume 9, pages 305–331, 1959.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 3rd edition, 1976.

Part I

Appendix

Table of Contents

A Additional Supporting lemmas and proofs **11**
A.1 Proof of Remark A.1 . . . 11
A.2 On existence of extreme points of $\mathcal{P}$ . . . 12
A.3 Lower and Upper probabilities are always extreme points . . . 13
A.4 Equivalence of extreme points for elicitation . . . 14
A.5 Preference relation in the subset of a credal set . . . 15

B Proof of Results in Section 3 **16**
B.1 Proof of Proposition 3.5 . . . 16
B.2 Proof of Proposition 3.6 . . . 16
B.3 Proof of Theorem 4.2 . . . 16

C Proof of Results in Section 4 **18**
C.1 Why is Non-dictatorship Desirable? . . . 18
C.2 Proof of Proposition 4.9 . . . 18
C.3 Proof of Lemma 4.10 . . . 19

D Proof for Results in Section 5 **21**
D.1 Proof of Lemma 5.3 . . . 21
D.2 Proof of Theorem 5.4 . . . 21

E Simulations **22**

A ADDITIONAL SUPPORTING LEMMAS AND PROOFS

A.1 PROOF OF REMARK A.1

Remark A.1. Scoring rule s is (strictly) proper if and only if the corresponding (strictly) convex function $G(q) = \sum_{o\in\mathcal{O}} s(q,o)q(o)$

Proof. It follows from Theorem 2.4 that regular scoring rule s is (strictly) proper if and only if there exists a corresponding (strictly) convex function G on $\Delta(\mathcal{O})$ such that

$$s(q,o) = G(q) - \sum_{o\in\mathcal{O}} G'(q)(o)q(o) + G'(q)(o). \tag{8}$$

($\Rightarrow$) Let us assume that there exists a strictly proper scoring rule s. Then, according to Theorem 2.4 there exists a convex function $G : \Delta(\mathcal{O}) \to \mathbb{R}$ such that

$$\begin{aligned} s(q,o) &= G(q) - \sum_{o\in\mathcal{O}} G'(q)(o)dq(o) + G'(q)(o) \\ \mathbb{E}_{o\sim p}[s(q,o)] &= \mathbb{E}_{o\sim p}\Big[G(q) - \sum_{o\in\mathcal{O}} G'(q)(o)dp(o) + G'(q)(o)\Big] && \text{(expert's true belief } p\text{)} \\ &= G(q) - \sum_{o\in\mathcal{O}} G'(q)(o)q(o) + \sum_{o\in\mathcal{O}} G'(q)(o)p(o), \end{aligned}$$

where q is the true belief of the forecaster. Then, we consider the maximum expected score

$$\begin{aligned}
\sum_{o\in\mathcal{O}} s(q,o)q(o) &:= \max_{p\in\Delta(\mathcal{O})} u_q(p) && (s \text{ is strictly proper})\\
&= \max_{p\in\Delta(\mathcal{O})} \mathbb{E}_{o\sim q}[s(p,o)]\\
&= \max_{p\in\Delta(\mathcal{O})} \left\{G(p) - \sum_{o\in\mathcal{O}} G'(p)(o)p(o) + \sum_{o\in\mathcal{O}} G'(p)(o)q(o)\right\}\\
&= G(p^*) - \sum_{o\in\mathcal{O}} G'(p^*)(o)p^*(o) + \sum_{o\in\mathcal{O}} G'(p^*)(o)q(o) && (p^* \text{ is the maximizer})\\
&= G(q) - \sum_{o\in\mathcal{O}} G'(q)(o)q(o) + \sum_{o\in\mathcal{O}} G'(q)(o)q(o) && (s \text{ is strictly proper so } p^* = q)\\
&= G(q).
\end{aligned}$$

($\Leftarrow$)

We define the (strictly) convex function G using the expected score of some scoring rule s, i.e., $G(p) = \sum_{o\in\mathcal{O}} s(p,o)p(o)$ and the subgradient $G'(p) = s(p,o)$. Then,

$$\begin{aligned}
G(p) - \sum_{o\in\mathcal{O}} G'(p)(o)p(o) + G'(p)(o) &:= \sum_{o\in\mathcal{O}} s(p,o)p(o) - \sum_{o\in\mathcal{O}} s(p,o)p(o) + s(p,o)\\
&= s(p,o)
\end{aligned}$$

This implies that s is a strictly proper scoring rule as a consequence of Theorem 2.4. □

A.2 ON EXISTENCE OF EXTREME POINTS OF $\mathcal{P}$

If the set of probability distributions $\mathcal{P}$ is infinite, the extreme points may not always exist. Therefore, we need to identify the conditions under which $\mathcal{P}$ has a valid set of extreme points. We argue that for extreme points to exist for $\mathcal{P}$, $\text{co}(\mathcal{P})$ must equal $\text{co}(\overline{\mathcal{P}})$, where $\overline{\mathcal{P}}$ is the closure of set $\mathcal{P}$.

Table 1: A toy example on existence of $\text{ext}(\mathcal{P})$ for $\mathcal{P}$ made using $p_1, p_2 \in \Delta(\mathcal{O})$

$\mathcal{P}$	Closed?	$ext(\mathcal{P})$ exists?	$\text{co}(\mathcal{P}) = \text{co}(\overline{\mathcal{P}})$
$\{p_1, p_2\}$	closed	yes	yes
$\{p \mid p := wp_1 + (1-w)p_2\ \forall w \in [0,1] - \{\frac{1}{2}\}\}$	open	yes	yes
$\{p \mid p := wp_1 + (1-w)p_2\ \forall w \in (0,1)\}$	open	no	no
$\{p \mid p := wp_1 + (1-w)p_2\ \forall w \in [0,1]\}$	closed	yes	yes

To show this we precisely define the extreme points of $\mathcal{P}$, independent of the convex hull of $\mathcal{P}$ as follows

Definition A.2. Given a set $\mathcal{P}$, we define the extreme points as $\text{ext}(\mathcal{P})$ as the collection of $p \in \mathcal{P}$ for which there does not exist a set of points $C \subseteq \mathcal{P} \setminus \{p\}$ and a probability measure $w : \Delta \to [0,1]$ such that $p = \int_C w(q)dq$.

For extreme points of a set to exist in general spaces, its convex hull must be compact according to Choquet's Theorem Bishop and Leeuw [1959]. To establish compactness of $\mathcal{P}$ we first show that with an appropriate notion of distance $\Delta(\mathcal{O})$ can form a bounded metric space.

Proposition A.3. *The metric space $(\Delta(\mathcal{O}), d_{TV})$ is bounded, where d_{TV} denotes the total variational distance between two probability distributions p, q in terms of their corresponding probability measures P, Q is defined as*

$$d_{TV}(p,q) := \sup_{A\subseteq\mathcal{O}} |P(A) - Q(A)| = \frac{1}{2}\int |p(o) - q(o)|do$$

Proof. As defined above, the total variational distance is half the L1 distance Levin and Peres [2017]. This allows us to express the total variation distance directly using densities. To show $(\Delta(\mathcal{O}), d_{TV})$ is bounded, let p, q be arbitrary distributions in $\Delta(\mathcal{O})$. Then

$$\begin{aligned}
d_{TV}(p,q) &:= \frac{1}{2}\int |p(o) - q(o)|do \\
&< \frac{1}{2}\int |p(o)| + |-q(o)|do && (\text{ Triangle Inequality}) \\
&= \frac{1}{2}\int |p(o)|do + \frac{1}{2}\int |q(o)|do \\
&= \frac{1}{2}\int p(o)do + \frac{1}{2}\int q(o)do && (p(o) \geq 0 \text{ and } q(o) \geq 0) \\
&= \frac{1}{2} + \frac{1}{2} = 1
\end{aligned}$$

Thus $(\Delta(\mathcal{O}), d_{TV})$ is a bounded metric space. □

We now discuss the conditions on $\mathcal{P}$ such that $\text{ext}(\mathcal{P}) \subseteq \mathcal{P}$.

Proposition A.4. *There exists a probability measure $w \in \Delta(ext(\mathcal{P}))$ for all $p \in \mathcal{P}$ such that*

$$p = \int_{p \in ext(\mathcal{P})} w(p)dp$$

iff $co(\mathcal{P}) = co(\overline{\mathcal{P}})$, where $co(\overline{\mathcal{P}})$ denotes the convex hull of the closure of $\mathcal{P}$ when $\mathcal{O}$ is finite. And for cases where $\mathcal{O}$ is an infinite continuous set, $co(\mathcal{P})$ must additionally be totally bounded.

Proof. The above result is a direct implication of the Heine-Borel Theorem (Theorem 2.41, [Rudin, 1976]) and Choquet's theorem [Bishop and Leeuw, 1959]. First we discuss the proof for the case where $\mathcal{O}$ is finite. Since $\mathcal{P} \subseteq \Delta(\mathcal{O})$, using A.3 we can say that $\mathcal{P}$ is bounded. This means that $\text{co}(\mathcal{P})$ is also bounded. Now, we know that the convex hull of a closed set $\overline{\mathcal{P}}$ is also closed. Therefore, $\text{co}(\overline{\mathcal{P}})$ is closed and since $\text{co}(\mathcal{P}) = \text{co}(\overline{\mathcal{P}})$, $\text{co}(\mathcal{P})$ is also closed. This makes $\text{co}(\mathcal{P})$ compact as it is both bounded and closed by Heine-Borel Theorem. Now we can directly apply Choquet's theorem to obtain a probability measure w for every $p \in \mathcal{P}$ such that $p = \int_{p \in \text{ext}(\mathcal{P})} w(p)dp$. In case when $\mathcal{O}$ is an infinite continuous set, we are dealing with $\mathcal{P} \subseteq \Delta(\mathcal{O})$, where $\Delta(\mathcal{O})$ may not have Heine-Borel Property, thus $\text{co}(\mathcal{P})$ being totally bounded in addition to closed ensures that $\text{co}(\mathcal{P})$ is compact and therefore Choquet's theorem is applicable. □

The proposition A.4 tries to identify what conditions should $\mathcal{P}$ satisfy so that we can interpret $\text{co}(\mathcal{P})$, i.e., the convex hull of $\mathcal{P}$ as a credal set with valid extreme points $\text{ext}(\mathcal{P})$. The general condition is that $\text{co}(\mathcal{P}) = \text{co}(\overline{\mathcal{P}})$ as a condition on $\mathcal{P}$. Equivalently, the condition on $\text{co}(\mathcal{P})$ is that it is closed. Notice that for finite $\mathcal{P}$, it is trivially satisfied. This allows us to exclude $\mathcal{P} = (0, \frac{1}{2})$ type of open sets from our discussion since they are open sets and its convex hull will violate closedness i.e. $\text{co}(\mathcal{P}) = \mathcal{P} = (0, \frac{1}{2})$. Depending on the convention, if credal sets for $\mathcal{P}$ are defined as the closure of their convex hulls, i.e., $\overline{\text{co}}(\mathcal{P})$, then credal sets are compact (Heine-Borel Theorem) and Proposition A.4 is applicable. Thus for our , we will restrict our discussion to $\mathcal{P}$ such that $\text{co}(\mathcal{P}) = \text{co}(\overline{\mathcal{P}})$.

A.3 LOWER AND UPPER PROBABILITIES ARE ALWAYS EXTREME POINTS

Lemma A.5. *Let $\mathcal{P} \subseteq \Delta(\mathcal{O})$ be the forecaster's belief and $ext(\mathcal{P})$ the extreme points of the convex hull generated by $\mathcal{P} \subseteq \Delta(\mathcal{O})$. Given any scoring rule $s : 2^{\Delta(\mathcal{O})} \times \mathcal{O} \to \mathbb{R}$ and forecasts $\mathcal{Q} \subseteq \Delta(\mathcal{O})$, let*

$$p_L^{(s,\mathcal{Q})} := \arg\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)], \qquad p_U^{(s,\mathcal{Q})} := \arg\sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)].$$

Then, both $p_L^{(s,\mathcal{Q})}$ and $p_U^{(s,\mathcal{Q})}$ belong to $ext(\mathcal{P})$ for all pairs of s and $\mathcal{Q}$. In addition, $\mathcal{P} \simeq ext(\mathcal{P})$.

Proof. Firstly, for all $p \in \mathcal{P}$, either $p \in \text{ext}(\mathcal{P})$ or $p \notin \text{ext}(\mathcal{P})$. This follows trivially from the definition of extreme points of a convex hull in section 2.2.

The proof proceeds as follows. In **(i)** and **(ii)**, we show that $p_{\text{L}}^{(s,\mathcal{Q})}, p_{\text{U}}^{(s,\mathcal{Q})} \in \text{ext}(\mathcal{P})$ for all pairs of s and $\mathcal{Q}$, respectively, with a contradiction. Then, given **(i)** and **(ii)**, $\mathcal{P} \simeq \text{ext}(\mathcal{P})$ follows from Definition 3.4 for the equivalence of imprecise beliefs.

(i) Lower probability: For all s and $\mathcal{Q}$, $p_{\text{L}}^{(s,\mathcal{Q})} \in \text{ext}(\mathcal{P})$.

We prove this by contradiction. Let us first assume there exists a pair of $s, \mathcal{Q}$ such that $p_{\text{L}}^{(s,\mathcal{Q})} \in \mathcal{P} \setminus \text{ext}(\mathcal{P})$. Since we have fixed s and $\mathcal{Q}$, we drop the superscript from $p_{\text{L}}^{(s,\mathcal{Q})}$ for readability and treat p_{L} as a distribution in $\mathcal{P}$. Next, given $p_{\text{L}} \in \mathcal{P} \setminus \text{ext}(\mathcal{P})$, it implies that there exists a second order distribution $w \in \Delta(\text{ext}(\mathcal{P}))$ such that $w(p) > 0$ for all $p \in \text{ext}(\mathcal{P})$.

$$
\begin{aligned}
p_{\text{L}} &= \int_{p \in \text{ext}(\mathcal{P})} w(p) dp \\
(\implies) \quad \mathbb{E}_{p_{\text{L}}}[s(\mathcal{Q}, o)] &= \mathbb{E}_{\int_{p \in \text{ext}(\mathcal{P})} w(p) dp}[s(\mathcal{Q}, o)] \\
&= \int_{p \in \text{ext}(\mathcal{P})} w(p) \mathbb{E}_p[s(\mathcal{Q}, o)] dp \\
&> \inf_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)]. \qquad (w(p) > 0 \text{ for all } p \in \text{ext}(\mathcal{P}))
\end{aligned}
$$

This results in a contradiction because $\text{ext}(\mathcal{P}) \subseteq \mathcal{P}$. Therefore, $p_{\text{L}} \in ext(\mathcal{P})$.Since our choice of $\mathcal{Q}$ and s was arbitrary, the contradiction holds for all $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ and s. Therefore, $p_{\text{L}}^{(s,\mathcal{Q})} \in \text{ext}(\mathcal{P})$ for all and s.

(ii) Upper probability: For all s and $\mathcal{Q}$, $p_{\text{U}}^{(s,\mathcal{Q})} \in \text{ext}(\mathcal{P})$.

Similarly, we show that $p_{\text{U}} \in \text{ext}(\mathcal{P})$. Suppose that $p_{\text{U}} \in \mathcal{P} \setminus \text{ext}(\mathcal{P})$. This implies that there exists a second order distribution $w \in \Delta(\text{ext}(\mathcal{P}))$ such that $w(p) > 0$ for all $p \in \text{ext}(\mathcal{P})$ and

$$
\begin{aligned}
p_{\text{U}} &= \int_{p \in \text{ext}(\mathcal{P})} w(p) dp \\
(\implies) \quad \mathbb{E}_{p_{\text{U}}}[s(\mathcal{Q}, o)] &= \mathbb{E}_{\int_{p \in \text{ext}(\mathcal{P})} w(p) dp}[s(\mathcal{Q}, o)] \\
&= \int_{p \in \text{ext}(\mathcal{P})} w(p) \mathbb{E}_p[s(\mathcal{Q}, o)] dp \\
&< \sup_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)]. \qquad (w(p) > 0 \text{ for all } p \in \text{ext}(\mathcal{P}))
\end{aligned}
$$

This also results in a contradiction since $\text{ext}(\mathcal{P}) \subseteq \mathcal{P}$. Hence, both p_{L} and p_{U} belong to $\text{ext}(\mathcal{P})$.

Equivalence of $\mathcal{P}$ and ext($\mathcal{P}$): Next, we show that $\mathcal{P}$ and $\text{ext}(\mathcal{P})$ are equivalent by applying Definition 3.4. For any reported set of beliefs $Q \subseteq \Delta(\mathcal{O})$ and scoring rule s,

$$
\begin{aligned}
\text{im}(\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})) &= \left[\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)],\ \sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] \right] \\
&= \left[\mathbb{E}_{p_{\text{L}}}[s(\mathcal{Q}, o)],\ \mathbb{E}_{p_{\text{U}}}[s(\mathcal{Q}, o)] \right] \\
&= \left[\inf_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)],\ \sup_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)] \right] \qquad (p_{\text{L}}, p_{\text{U}} \in \text{ext}(\mathcal{P}) \text{ and } \text{ext}(\mathcal{P}) \subseteq \mathcal{P}) \\
&= \text{im}(\boldsymbol{V}^{\text{ext}(\mathcal{P})}(\mathcal{Q})).
\end{aligned}
$$

This completes the proof. □

A.4 EQUIVALENCE OF EXTREME POINTS FOR ELICITATION

Lemma A.6. *If two beliefs $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{O})$ are equivalent, i.e., $\mathcal{P} \simeq \mathcal{P}'$, then ext($\mathcal{P}$) = ext($\mathcal{P}'$).*

Proof. By Definition 3.4, two imprecise beliefs $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{O})$ are equivalent if for all scoring rule s and forecast $\mathcal{Q} \subseteq \Delta(\mathcal{O})$, $\text{im}(\boldsymbol{V}^{\mathcal{P}}(\mathcal{Q})) = \text{im}(\boldsymbol{V}^{\mathcal{P}'}(Q))$. This means that,

$$\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] = \inf_{p' \in \mathcal{P}'} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] \quad \text{and} \quad \sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] = \sup_{p' \in \mathcal{P}'} \mathbb{E}_{p'}[s(\mathcal{Q}, o)].$$

($\Rightarrow$) For the first part of the proof, we show that $\text{ext}(\mathcal{P}) \subseteq \text{ext}(\mathcal{P}')$. Let $q \in \text{ext}(\mathcal{P})$, we know that for all $s, \mathcal{Q}$,

$$\begin{aligned}
\inf_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)] \leq \mathbb{E}_q[s(\mathcal{Q}, o)] &\leq \sup_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)] \\
\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] \leq \mathbb{E}_q[s(\mathcal{Q}, o)] &\leq \sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] && (\mathcal{P} \simeq \text{ext}(\mathcal{P}) \text{ from Lemma } A.5) \\
\inf_{p' \in \mathcal{P}'} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] \leq \mathbb{E}_q[s(\mathcal{Q}, o)] &\leq \sup_{p' \in \mathcal{P}'} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] && (\mathcal{P} \simeq \mathcal{P}' \text{ by definition}) \\
\inf_{p' \in \text{ext}(\mathcal{P}')} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] \leq \mathbb{E}_q[s(\mathcal{Q}, o)] &\leq \sup_{p' \in \text{ext}(\mathcal{P}')} \mathbb{E}_{p'}[s(\mathcal{Q}, o)]. && (\mathcal{P}' \simeq \text{ext}(\mathcal{P}') \text{ from Lemma } A.5)
\end{aligned}$$

The last inequalities imply that $q \in \text{ext}(\mathcal{P}')$. ($\Leftarrow$) Next, we show that $\text{ext}(\mathcal{P}') \subseteq \text{ext}(\mathcal{P})$. Let $q' \in \text{ext}(\mathcal{P}')$. Then, we know that for all $s, \mathcal{Q}$,

$$\begin{aligned}
\inf_{p' \in \text{ext}(\mathcal{P}')} \mathbb{E}_p[s(\mathcal{Q}, o)] \leq \mathbb{E}_{q'}[s(\mathcal{Q}, o)] &\leq \sup_{p' \in \text{ext}(\mathcal{P}')} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] \\
\inf_{p' \in \mathcal{P}'} \mathbb{E}_{p'}[s(\mathcal{Q}, o)] \leq \mathbb{E}_{q'}[s(\mathcal{Q}, o)] &\leq \sup_{p' \in \mathcal{P}'} \mathbb{E}_p[s(\mathcal{Q}, o)] && (\mathcal{P}' \simeq \text{ext}(\mathcal{P}') \text{ from Lemma } A.5) \\
\inf_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] \leq \mathbb{E}_{q'}[s(\mathcal{Q}, o)] &\leq \sup_{p \in \mathcal{P}} \mathbb{E}_p[s(\mathcal{Q}, o)] && (\mathcal{P} \simeq \mathcal{P}' \text{ by definition}) \\
\inf_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)] \leq \mathbb{E}_{q'}[s(\mathcal{Q}, o)] &\leq \sup_{p \in \text{ext}(\mathcal{P})} \mathbb{E}_p[s(\mathcal{Q}, o)]. && (\mathcal{P} \simeq \text{ext}(\mathcal{P}) \text{ from Lemma } A.5)
\end{aligned}$$

The last inequalities imply that $q' \in \text{ext}(\mathcal{P})$. Since both $\text{ext}(\mathcal{P}) \subseteq \text{ext}(\mathcal{P}')$ and $\text{ext}(\mathcal{P}') \subseteq \text{ext}(\mathcal{P})$, we can conclude that $\text{ext}(\mathcal{P}) = \text{ext}(\mathcal{P}')$. □

A.5 PREFERENCE RELATION IN THE SUBSET OF A CREDAL SET

The lemma argues that the dominance induced by the preference relation associated with a credal set can only be refined by considering its subsets. Formally,

Lemma A.7. *For any pair of imprecise forecasts $\mathcal{P}, \mathcal{Q} \subseteq \Delta(\mathcal{O})$ such that $co(\mathcal{Q}) \subset co(\mathcal{P})$*

$$a \succeq_{\mathcal{P}} a' \implies a \succeq_{\mathcal{Q}} a' \quad \forall\, a, a' \in \mathcal{A}$$

where $\succeq_{\mathcal{P}}, \succeq_{\mathcal{Q}}$ are the partial preference relations over the space of actions induced by the corresponding expected utility profiles $\{\mathbb{E}_p[u(\cdot, o)]\}_{p \in \mathcal{P}}$ and $\{\mathbb{E}_q[u(\cdot, o)]\}_{q \in \mathcal{Q}}$.

Proof. Let us assume an arbitrary $\mathcal{Q}$ and $\mathcal{P}$ such that $\mathcal{Q} \subset \mathcal{P}$. Now let us consider a pair of inputs $x, x' \in \mathcal{X}$ such that $x \succeq_{\mathcal{P}} x'$. This implies that

$$\begin{aligned}
& \mathbb{E}_p[u(x, o)] \geq \mathbb{E}_p[u(x', o)] \quad \forall p \in \mathcal{P} \\
\implies\ & \mathbb{E}_p[u(x, o)] \geq \mathbb{E}_p[u(x', o)] \quad \forall p \in co(\mathcal{P}) \\
\implies\ & \mathbb{E}_q[u(x, o)] \geq \mathbb{E}_q[u(x', o)] \quad \forall q \in co(\mathcal{Q}) && (co(\mathcal{Q}) \subset co(\mathcal{P})) \\
\implies\ & \mathbb{E}_q[u(x, o)] \geq \mathbb{E}_q[u(x', o)] \quad \forall q \in \mathcal{Q} \\
& \implies \quad x \succeq_{\mathcal{Q}} x'
\end{aligned}$$

□

B PROOF OF RESULTS IN SECTION 3

B.1 PROOF OF PROPOSITION 3.5

Proof. ($\Leftarrow$) Let us assume that there are two identical distributions $p, q \in \Delta(\mathcal{O})$, i.e., $p = q$ that implies $\mathbb{E}_p[s(\mathcal{Q}, o)] = \mathbb{E}_q[s(\mathcal{Q}, o)]$ for all $\mathcal{Q} \subseteq \Delta(\mathcal{O})$ and IP scoring rule s. Therefore, $\{p\} \simeq \{q\}$.

($\Rightarrow$) Next, let us assume that $\{p\} \simeq \{q\}$, which means that

$$\mathbb{E}_p[s(\mathcal{Q}, o)] = \mathbb{E}_q[s(\mathcal{Q}, o)], \quad \forall s, \forall \mathcal{Q} \subseteq \Delta(\mathcal{O}).$$

Since the above holds for all s and $\mathcal{Q}$, we choose s to be strictly proper for precise forecasts and $\mathcal{Q} := \{p\}$. Hence,

$$\begin{aligned} &\mathbb{E}_p[s(\{p\}, o)] = \mathbb{E}_q[s(\{p\}, o)] \\ &\implies \quad p = q. && (s \text{ is strictly proper for precise forecasts}) \end{aligned}$$

This completes the proof. □

B.2 PROOF OF PROPOSITION 3.6

Proposition B.1. *For imprecise beliefs $\mathcal{P}, \mathcal{P}' \subseteq \Delta(\mathcal{O})$ with non-empty extreme points, $\mathcal{P} \simeq \mathcal{P}'$ if and only if $co(\mathcal{P}) = co(\mathcal{P}')$.*

Proof. ($\Leftarrow$) First, we assume that $\mathcal{P}$ and $\mathcal{P}'$ induce the same credal set, i.e.,

$$\begin{aligned} \text{co}(\mathcal{P}) &= \text{co}(\mathcal{P}') \\ \text{ext}(\mathcal{P}) &= \text{ext}(\mathcal{P}') && \text{(credal sets are convex hulls)} \\ \text{ext}(\mathcal{P}) \simeq \mathcal{P} \quad &\text{and} \quad \text{ext}(\mathcal{P}') \simeq \mathcal{P}' && \text{(Lemma A.5)} \\ \implies \quad \mathcal{P} &\simeq \mathcal{P}' \end{aligned}$$

Hence, $\mathcal{P}$ and $\mathcal{P}'$ are equivalent.

($\Rightarrow$) Next, we assume that $\mathcal{P}$ and $\mathcal{P}'$ are equivalent. Then, it follows from Lemma A.6 that $\text{ext}(\mathcal{P}) = \text{ext}(\mathcal{P}')$.

Let us assume that there exists a $P \in \text{co}(\mathcal{P})$. Since credal sets are convex sets, P can be expressed as a convex combination of the extreme points. Therefore, there exists some $w \in \Delta(\text{ext}(\mathcal{P}))$ such that

$$P = \int_{p \in \text{ext}\mathcal{P}} w(p)dp \underset{(\blacklozenge)}{=} \int_{p \in \text{ext}\mathcal{P}'} w(p)dp \qquad (\blacklozenge : \text{Lemma A.6})$$

Thus $P \in \text{co}(\mathcal{P}')$ and therefore, $\text{co}(\mathcal{P}') \subseteq \text{co}(\mathcal{P})$.

Similarly, let us assume that there exists a $P' \in \text{co}(\mathcal{P}')$. Now P' can also be expressed as a convex combination of the extreme points. Therefore, there exists some $w' \in \Delta(\text{ext}(\mathcal{P}'))$ such that

$$P' = \int_{p' \in \text{ext}\mathcal{P}'} w'(p')dp' \underset{(\blacklozenge)}{=} \int_{p' \in \text{ext}\mathcal{P}} w(p')dp' \qquad (\blacklozenge : \text{Lemma A.6})$$

Thus $P \in \text{co}(\mathcal{P})$ and therefore, $\text{co}(\mathcal{P}) \subseteq \text{co}(\mathcal{P}')$. Since $\text{co}(\mathcal{P}') \subseteq \text{co}(\mathcal{P})$ and $\text{co}(\mathcal{P}) \subseteq \text{co}(\mathcal{P}')$, therefore $\text{co}(\mathcal{P}) = \text{co}(\mathcal{P}')$ □

B.3 PROOF OF THEOREM 4.2

Part I: We first show that for any IP scoring rule s, it must give a constant score to all forecasts.

Proof. Let us assume there exists a proper scoring rule $s : 2^{\Delta(\mathcal{O})} \times \mathcal{O} \to \mathbb{R} \cup \{-\infty\}$. Then, according to the definition of proper IP scoring rules for an imprecise forecaster with a vacuous belief $\Delta(\mathcal{O})$, we must have,

$$\Delta(\mathcal{O}) \succeq_{\Delta(\mathcal{O})} \mathcal{Q}, \quad \forall \mathcal{Q} \neq \Delta(\mathcal{O}).$$

This follows from the fact that for a proper score s, $\boldsymbol{V}^{\Delta(\mathcal{O})}(\Delta(\mathcal{O}))$ dominates $\boldsymbol{V}^{\Delta(\mathcal{O})}(\mathcal{Q})$. Consequently, it follows from Definition 4.1 that

$$\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] \geq \mathbb{E}_p[s(\mathcal{Q}, o)], \quad \forall \mathcal{Q} \neq \Delta(\mathcal{O}),\ \forall p \in \Delta(\mathcal{O}). \tag{9}$$

Let $\tilde{\mathcal{Q}} := \{\mathcal{Q} \mid \mathcal{Q} \neq \Delta(\mathcal{O})\}$ be the set of all forecasts not equivalent to the forecaster's belief ($\Delta(\mathcal{O})$), then we can rewrite (9) as

$$\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] \geq \mathbb{E}_p[s(\mathcal{Q}, o)], \quad \forall \mathcal{Q} \in \tilde{\mathcal{Q}},\ \forall p \in \Delta(\mathcal{O}). \tag{10}$$

Also, $\{q\}_{q \in \Delta(\mathcal{O})} \subseteq \tilde{\mathcal{Q}}$ since $q \neq \Delta(\mathcal{O})$. Combining this with Equation (10) yields

$$\begin{aligned} & \mathbb{E}_p[s(\Delta(\mathcal{O}), o)] \geq \mathbb{E}_p[s(\{q\}, o)], \quad \forall q \in \Delta(\mathcal{O}),\ \forall p \in \Delta(\mathcal{O}) \\ \implies \quad & \mathbb{E}_p[s(\Delta(\mathcal{O}), o)] \geq \mathbb{E}_p[s(\{p\}, o)], \quad \forall p \in \Delta(\mathcal{O}), \end{aligned} \tag{11}$$

where the second inequalities follow by selecting the inequalities such that $q = p$. Similarly, let us analyse the incentives for all precise forecasters with belief $p \in \Delta(\mathcal{O})$ given a proper IP scoring rule s. Then, for all precise forecasters we must have,

$$\begin{aligned} & \{p\} \succeq_{\{p\}} \mathcal{Q}, \quad \forall p \in \Delta(\mathcal{O}),\ \forall \mathcal{Q} \in \tilde{\mathcal{Q}} && (\tilde{\mathcal{Q}} := 2^{\Delta(\mathcal{O})} \setminus \{p\}) \\ \implies \quad & \{p\} \succeq_{\{p\}} \Delta(\mathcal{O}), \quad \forall p \in \Delta(\mathcal{O}) && (\Delta(\mathcal{O}) \in \tilde{\mathcal{Q}}) \\ \implies \quad & \mathbb{E}_p[s(\{p\}, o)] \geq \mathbb{E}_p[s(\Delta(\mathcal{O}), o)], \quad \forall p \in \Delta(\mathcal{O}). \end{aligned}$$

However, it follows from Equation (11) that $\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] \geq \mathbb{E}_p[s(\{p\}, o)]$ and $\mathbb{E}_p[s(\{p\}, o)] \geq \mathbb{E}_p[s(\Delta(\mathcal{O}), o)]$ for all $p \in \Delta(\mathcal{O})$. This implies that $\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] = \mathbb{E}_p[s(\{p\}, o)]$ for all $p \in \Delta(\mathcal{O})$.

Therefore, any IP scoring rule s that satisfies properness sets up incorrect incentives for the forecaster. For example, the expected score for honestly reporting a precise forecast is the same as reporting the vacuous set of all distributions, i.e.,

$$\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] = \mathbb{E}_p[s(\{p\}, o)], \quad \forall p \in \Delta(\mathcal{O}). \tag{12}$$

While the above equation is sufficient to discard any proper scoring rule, we show that the only IP scoring rule possible is a constant function. For s to be proper for imprecise forecasts, the following must hold true for all $\mathcal{P} \subseteq \Delta(\mathcal{O})$:

$$\begin{aligned} & \mathcal{P} \succeq_{\mathcal{P}} \{q\}, \quad \forall q \in \Delta(\mathcal{O}) \\ & \mathbb{E}_p[s(\mathcal{P}, o)] \geq \mathbb{E}_p[s(\{q\}, o)], \quad \forall q \in \Delta(\mathcal{O}),\ \forall p \in \mathcal{P} \\ & \mathbb{E}_p[s(\mathcal{P}, o)] \geq \mathbb{E}_p[s(\{q\}, o)], \quad \forall q \in \mathcal{P},\ \forall p \in \mathcal{P} \\ \implies & \mathbb{E}_p[s(\mathcal{P}, o)] \geq \mathbb{E}_p[s(\{p\}, o)], \quad \forall p \in \mathcal{P}. \end{aligned} \tag{13}$$

Similarly, for any $p \in \Delta(\mathcal{O})$, the following must hold:

$$\begin{aligned} & \{p\} \succeq_p \mathcal{P}, \quad \forall p \in \Delta(\mathcal{O}) \\ \implies & \mathbb{E}_p[s(\{p\}, o)] \geq \mathbb{E}_p[s(\mathcal{P}, o)]. \end{aligned} \tag{14}$$

Combining Equations 12, 13 and 14 yields

$$\mathbb{E}_p[s(\Delta(\mathcal{O}), o)] = \mathbb{E}_p[s(\{p\}, o)] = \mathbb{E}_p[s(\mathcal{P}, o)], \quad \forall p \in \Delta(\mathcal{O}) \tag{15}$$

Given Equation 15 is valid for all $p \in \Delta(\mathcal{O})$, we consider the a subset of $\Delta(\mathcal{O})$. To be precise, the set of all Dirac distributions associated with each outcome, i.e. $p \in \{\delta_o\}_{o \in \mathcal{O}}$

$$\begin{aligned} & \mathbb{E}_p[s(\Delta(\mathcal{O}), o)] = \mathbb{E}_p[s(\{p\}, o)] = \mathbb{E}_p[s(\mathcal{P}, o)], \quad p \in \{\delta_o\}_{o \in \mathcal{O}} && (\{\delta_o\}_{o \in \mathcal{O}} \subseteq \Delta(\mathcal{O})) \\ \implies & s(\Delta(\mathcal{O}), o) = s(\{p\}, o) = s(\mathcal{P}, o), \quad \forall o \in \mathcal{O}. \end{aligned}$$

Hence, s needs to be a constant score for it to be a proper IP scoring rule. □

Part II: There exists no strictly proper IP scoring rule s.

Proof. Assume that there exists a strictly proper IP scoring rule s. Consider a precise forecaster with belief $q \in \Delta(\mathcal{O})$. Then, we have

$$\begin{aligned} &\{q\} \succ_q \mathcal{Q}, \quad \forall \mathcal{Q} \neq q \\ \implies &\mathbb{E}_q[s(\{q\}, o)] > \mathbb{E}_q[s(\mathcal{Q}, o)] \\ \implies &\mathbb{E}_q[s(\{q\}, o)] > \mathbb{E}_q[s(\Delta(\mathcal{O}), o)]. \qquad (\Delta(\mathcal{O}) \text{ is one possible } \mathcal{Q}) \end{aligned} \tag{16}$$

Since s is strictly proper, it satisfies Equation 12. However, this results in a contradiction to Equation 16. Hence, no s can be strictly proper. □

C PROOF OF RESULTS IN SECTION 4

C.1 WHY IS NON-DICTATORSHIP DESIRABLE?

Let us assume that ρ violates non-dictatorship, then ρ is dictatorial. For clarity, we also define a dictatorship.

Definition C.1. (Dictatorship) An aggregation rule ρ is a dictatorial if there exists a $P_\rho \in \mathcal{P}$ (dictator), that depends on ρ, such that for any pair of reports $\mathcal{Q}, \mathcal{Q}' \subseteq \Delta(\mathcal{O})$,

$$\mathcal{Q} \succeq_{P_\rho} \mathcal{Q}' \implies \mathcal{Q} \succeq_{\rho[\mathcal{P}]} \mathcal{Q}'.$$

A dictatorial ρ not only allows the forecaster to remove indecision in their decision-making problem about which $\mathcal{Q}$ to report, it also allows the forecaster to precisely resolve their epistemic uncertainty, i.e., by reducing the credal set $\mathcal{P}$ to only the dictator P_ρ.

Let us denote the set of best reports plausible under aggregation ρ by $\tilde{\mathcal{Q}}^\rho := \{\mathcal{Q} \mid \mathcal{Q} \succeq_\rho \mathcal{Q}', \forall \mathcal{Q}' \subseteq \Delta(\mathcal{O})\}$. Since $\succeq_\rho$ is complete, if the set of best reports $\tilde{\mathcal{Q}}$ contains more than one report, then they must be indifferent w.r.t. $\succeq_{\rho[\mathcal{P}]}$. Given ρ is a dictatorship, there exists $P_\rho \in \mathcal{P}$ such that $\succeq_{P_\rho}$ dictates the preference $\succeq_{\rho[\mathcal{P}]}$. That is, the set of best reports under P_ρ must be exactly the same as that under ρ. Therefore,

$$\tilde{\mathcal{Q}}^P = \tilde{\mathcal{Q}}^\rho.$$

This implies that the expected scores of P_ρ and $\mathcal{P}$ with any dictatorial ρ is the same, i.e.,

$$V_\rho^{\mathcal{P}}(\{P_\rho\}) = V_\rho^{\mathcal{P}}(\mathcal{P}).$$

C.2 PROOF OF PROPOSITION 4.9

Proof. We prove this result by contradiction. Let us assume that there exists a tailored scoring rule s_ρ that is not proper and analyse this scoring rule for an arbitrary forecaster with an imprecise belief $\mathcal{P} \subseteq \Delta(\mathcal{O})$. Since s_ρ is not proper, it implies that there exists $Q \subseteq \Delta(\mathcal{O})$ where $\mathcal{Q} \neq \mathcal{P}$ such that

$$V_\rho^{\mathcal{P}}(\mathcal{Q}) > V_\rho^{\mathcal{P}}(\mathcal{P}). \tag{17}$$

In other words, the forecaster strictly prefers the forecast $\mathcal{Q}$ over their belief $\mathcal{P}$. However, let's analyse the scenario from DM's perspective when they obtain forecast $\mathcal{P}$, the optimal action according to the forecast $\mathcal{P}$ is

$$a^*_{\mathcal{P},\rho} = \arg\max_{a \in \mathcal{A}} \ \rho\left[\{\mathbb{E}_p[u(a, o)]\}_{p \in \mathcal{P}}\right]. \tag{18}$$

Since $a^*_{\mathcal{P},\rho}$ is the maximizer of DM's aggregated utility, this means that for all $a \in \mathcal{A}$,

$$\rho\left[\{\mathbb{E}_p[u(a^*_{\mathcal{P},\rho}, o)]\}_{p \in \mathcal{P}}\right] \geq \rho\left[\{\mathbb{E}_p[u(a, o)]\}_{p \in \mathcal{P}}\right]. \tag{19}$$

However, we know that from Equation (17)

$$\begin{aligned} V_\rho^{\mathcal{P}}(\mathcal{Q}) &> V_\rho^{\mathcal{P}}(\mathcal{P}) \\ \rho[\{\mathbb{E}_p[s_\rho(\mathcal{Q}, o)]\}_{p\in\mathcal{P}}] &> \rho[\{\mathbb{E}_p[s_\rho(\mathcal{P}, o)]\}_{p\in\mathcal{P}}] \\ \rho[\{\mathbb{E}_p[u(a^*_{\mathcal{Q},\rho}, o)]\}_{p\in\mathcal{P}}] &> \rho[\{\mathbb{E}_p[u(a^*_{\mathcal{P},\rho}, o)]\}_{p\in\mathcal{P}}]. \end{aligned}$$

This results in a contradiction to Equation (19). Therefore, s_ρ must be proper. Since this holds for any choice of ρ, we can conclude that s_ρ must be proper for any aggregation rule ρ. □

C.3 PROOF OF LEMMA 4.10

Part I: Strict properness of IP scoring rule for precise forecasts

Proof. (⟹) From Theorem 2.4, a regular precise scoring rule s is (strictly) proper if and only if there exists a corresponding (strictly) convex function G on $\Delta(\mathcal{O})$ such that

$$s(p, o) = G(p) - \sum_{o\in\mathcal{O}} G'(p)(o)p(o) + G'(p)(o).$$

Moreover, it follows from Remark A.1 that the $G(p) = \mathbb{E}_p[s(p, o)]$. Hence, for a tailored scoring rule s_ρ on precise distribution $p \in \Delta(\mathcal{O})$ to be strictly proper, we must have

$$\begin{aligned} G(p) &= \mathbb{E}_p[s_\rho(\{p\}, o)] \\ &= k\mathbb{E}_p[u(a^*_p, o)] + c && \text{(Tailored scoring rule; Definition 4.8)} \\ &= \max_{a\in\mathcal{A}} k\mathbb{E}_p[u(a, o)] + c. \end{aligned}$$

Next, for $G(p)$ to be strictly convex in p, we must have that for all $p, q \in \Delta(\mathcal{O})$,

$$G(q) > G(p) + \sum_{o\in\mathcal{O}} G'(p)(o)[q(o) - p(o)] \tag{20}$$

Where $G'(p)(o)$ is the oth component of the gradient $G'(p)$ at p. Let us consider the right-hand side of Equation (20).

$$\begin{aligned} G(p) + \sum_{o\in\mathcal{O}} G'(p)(o)[q(o) - p(o)] &= k\mathbb{E}_p[u(a^*_p, o)] + c + \sum_{o\in\mathcal{O}} ku(a^*_p, o)[q(o) - p(o)] && (21) \\ &= k\mathbb{E}_p[u(a^*_p, o)] + k\mathbb{E}_q[u(a^*_p, o)] - k\mathbb{E}_p[u(a^*_p, o)] + c && (22) \\ &= k\mathbb{E}_q[u(a^*_p, o)] + c. && (23) \end{aligned}$$

Since $G(q) := k\mathbb{E}_q[u(a^*_q, o)] + c$, for G to be strictly convex, we use Equation (23) to rewrite Equation (20) as follows

$$\begin{aligned} & G(q) > k\mathbb{E}_q[u(a^*_p, o)] + c, \quad \forall p, q \in \Delta(\mathcal{O}), \\ \implies \quad & k\mathbb{E}_q[u(a^*_q, o)] + c > k\mathbb{E}_q[u(a^*_p, o)] + c, \quad \forall p, q \in \Delta(\mathcal{O}). \end{aligned}$$

Hence, a^*_q must be a unique maximizer.

(⟸)

We assume that $a^*_p := \arg\max_{a\in\mathcal{A}} \mathbb{E}_p[u(a, o)]$ is the unique maximizer for all $p \in \Delta(\mathcal{O})$. Then, for all $p, q \in \Delta(\mathcal{O})$ and some arbitrary $\lambda \in [0, 1]$,

$$\begin{aligned} G(\lambda p + (1-\lambda)q) &= \mathbb{E}_{\lambda p+(1-\lambda)q}[s_\rho(\{\lambda p + (1-\lambda)q\}, o)] \\ &= \lambda\mathbb{E}_p[s_\rho(\{\lambda p + (1-\lambda)q\}, o)] + (1-\lambda)\mathbb{E}_q[s_\rho(\{\lambda p + (1-\lambda)q\}, o)] \\ &= \lambda k\mathbb{E}_p[u(a^*_{\lambda p+(1-\lambda)q}, o)] + \lambda c + (1-\lambda)k\mathbb{E}_q[u(a^*_{\lambda p+(1-\lambda)q}, o)] + (1-\lambda)c \\ &< \lambda k\mathbb{E}_p[u(a^*_p, o)] + \lambda c + (1-\lambda)k\mathbb{E}_q[u(a^*_q, o)] + (1-\lambda)c && (a^*_p \text{ and } a^*_q \text{ are unique}) \\ &= \lambda G(p) + (1-\lambda)G(q). \end{aligned}$$

Hence, G is strictly convex. □

Part II: Impossibility of strictly proper scoring rules with Pareto efficient ρ

Proof. Suppose that there exists the aggregation rule ρ such that the tailored scoring rule s_ρ is strictly proper for both precise and imprecise forecasts. This means that for all $\mathcal{P} \subseteq \Delta(\mathcal{O})$, and for all $\mathcal{Q} \nsubseteq \mathcal{P}$,

$$\begin{aligned} & V_\rho^{\mathcal{P}}(\mathcal{P}) > V_\rho^{\mathcal{P}}(\mathcal{Q}) \\ \implies \quad & \rho(\{\mathbb{E}_p[s_\rho(\mathcal{P}, o)]\}_{p\in\mathcal{P}}) > \rho(\{\mathbb{E}_p[s_\rho(\mathcal{Q}, o)]\}_{p\in\mathcal{P}}) \\ \implies \quad & \rho(\{\mathbb{E}_p[u(a^*_{\rho,\mathcal{P}}, o)]\}_{p\in\mathcal{P}}) > \rho(\{\mathbb{E}_p[u(a^*_{\rho,\mathcal{Q}}, o)]\}_{p\in\mathcal{P}}). \qquad (s_\rho \text{ is tailored scoring rule}) \end{aligned}$$

The aggregation rule ρ maps the set of preferences $\succeq_{\mathcal{P}} := \{\succeq_p\}_{p\in\mathcal{P}}$ into a complete preference relation $\succeq_{\rho(\mathcal{P})}$ which follows the aggregated utility $\rho(\{\mathbb{E}_p[u(\cdot, o)]\}_{p\in\mathcal{P}})$.

Since ρ is Pareto efficient, for all $a, a' \in \mathcal{A}$, $a \succeq_{\mathcal{P}} a'$ implies $a \succeq_{\rho[\mathcal{P}]} a'$. Only for actions that are incomparable to one another, i.e., $a \not\succeq_{\mathcal{P}} a'$ and $a' \not\succeq_{\mathcal{P}} a$, ρ decides to remove indecision by completing the preference as $a \succeq_{\rho[\mathcal{P}]} a'$ or $a' \succeq_{\rho[\mathcal{P}]} a$.

Without loss of generality, let us assume that ρ chooses to rank $a \succeq_{\rho[\mathcal{P}]} a'$ for two incomparable $a, a' \in \mathcal{A}$ with respect to original credal set $\mathcal{P}$. However, based on Lemma A.7, we can construct a $Q \subseteq \Delta(\mathcal{O})$ such that $\text{co}(\mathcal{Q}) \subset \text{co}(\mathcal{P})$ and $a^*_{\rho,\mathcal{P}} = a^*_{\rho,\mathcal{Q}}$. This provides a counterexample to strictness of s_ρ for all Pareto efficient ρ.

We now explain the counterexample in detail. We construct $\mathcal{Q}$ based on its partial preference relation $\succeq_Q$. The preference relation $\succeq_Q$ must be well defined for any two pair of actions $a, a' \in \mathcal{A}$. To this end we use the preference relation $\succeq_{\mathcal{P}}$ to define all possible scenarios for a pair of actions $a, a' \in \mathcal{A}$. Either $a, a' \in \mathcal{A}$ are comparable with respect to $\succeq_{\mathcal{P}}$ (**Case I**) or incomparable (**Case II**). The construction of $\succeq_{\mathcal{Q}}$ is defined below

Case I

$a \succeq_{\mathcal{P}} a'$ implies $a \succeq_{\mathcal{Q}} a'$

(Lemma A.7)

Case II

$(a \not\succeq_{\mathcal{P}} a') \wedge (a' \not\succeq_{\mathcal{P}} a)$

Case II.1

$a \succeq_{\mathcal{Q}} a' := a \succeq_{\rho[\mathcal{P}]} a'$

Case II.2

$(a \not\succeq_{\mathcal{Q}} a') \wedge (a' \not\succeq_{\mathcal{Q}} a)$

Figure 2: In **Case I** when actions are comparable in $\succeq_{\mathcal{P}}$ the Lemma A.7 dictates their order to be the same for partial preference induced by $\succeq_{\mathcal{Q}}$. However, in **Case II** when the actions are incomparable w.r.t $\succeq_{\mathcal{P}}$ either their order in $\succeq_{\mathcal{Q}}$ must be set to aggregated order of $\mathcal{P}$ i.e. $\succeq_{\rho[\mathcal{P}]}$ or they are left untouched, i.e. incomparable w.r.t $\succeq_{\mathcal{Q}}$

Now we are ready to reason what happens when we aggregate the partial preference $\succeq_{\mathcal{Q}}$ with ρ. We will reason for all the cases we defined above.

Case I: For all pairs of $a, a' \in \mathcal{A}$ that are comparable w.r.t. $\succeq_{\mathcal{P}}$ (Assume w.l.o.g $a \succeq_{\mathcal{P}} a'$).

$$a \succeq_{\mathcal{P}} a' \underset{(\clubsuit)}{\implies} a \succeq_{\rho[\mathcal{P}]} a' \quad \text{and} \quad a \succeq_{\mathcal{P}} a' \underset{(\blacklozenge)}{\implies} a \succeq_{\mathcal{Q}} a' \underset{(\clubsuit)}{\implies} a \succeq_{\rho[\mathcal{Q}]} a'. \qquad (\clubsuit : \rho \text{ is PE}, \blacklozenge\text{: Lemma A.7})$$

Therefore, whenever the pair of actions $a, a' \in \mathcal{A}$ are comparable w.r.t. $\succeq_{\mathcal{P}}$, the aggregated preference relation is the same, i.e., $\{\succeq_{\rho[\mathcal{P}]}\} \equiv \{\succeq_{\rho[\mathcal{Q}]}\}$.

Case II: Consider $a, a' \in \mathcal{A}$ that are incomparable w.r.t. $\succeq_{\mathcal{P}}$. (Assume w.l.o.g that ρ resolves this as $a \succeq_{\rho[\mathcal{P}]} a'$)

Case II.1: The pair of $a, a' \in \mathcal{A}$ is also comparable w.r.t. $\succeq_{\mathcal{Q}}$ (Assume w.l.o.g $a \succeq_{\mathcal{Q}} a'$)

$$a \succeq_{\mathcal{Q}} a' \underset{(\clubsuit)}{\implies} a \succeq_{\rho[\mathcal{Q}]} a' \quad \text{and} \quad a \succeq_{\mathcal{Q}} a' \underset{(\blacklozenge)}{\implies} a \succeq_{\rho[\mathcal{P}]} a' \qquad (\clubsuit : \rho \text{ is PE}, \blacklozenge : \text{by construction})$$

Case II.2: The pair $a, a' \in \mathcal{A}$ is incomparable w.r.t. $\succeq_{\mathcal{Q}}$.
Since ρ is a function, it will resolve indecision for two inputs in the same way, given that $|\mathcal{A}|$ is fixed across both these resolutions:

$$\Big((a \not\succeq_{\mathcal{P}} a' \wedge a' \not\succeq_{\mathcal{P}} a) \implies (a \succeq_{\rho[\mathcal{P}]} a')\Big) \wedge (a \not\succeq_{\mathcal{Q}} a' \wedge a' \not\succeq_{\mathcal{Q}} a) \implies a \succeq_{\rho[\mathcal{Q}]} a'.$$

Therefore, similar to Case 1, whenever the pair of actions $a, a' \in \mathcal{A}$ are incomparable w.r.t. $\succeq_{\mathcal{P}}$, the aggregated preference is the same, i.e., $\{\succeq_{\rho[\mathcal{P}]}\} \equiv \{\succeq_{\rho[\mathcal{Q}]}\}$. Hence, $a^*_{\rho,\mathcal{P}} = a^*_{\rho,\mathcal{Q}}$.

This makes s_ρ not strictly proper. □

D PROOF FOR RESULTS IN SECTION 5

D.1 PROOF OF LEMMA 5.3

Proof. We prove this by contradiction, let us assume that, $\mathcal{A}_{\text{ext}}$ is not a sufficient way to represent credal sets in the actions space. This implies that there exists a pair of credal sets $\mathcal{Q}, \mathcal{Q}' \subseteq \mathcal{O}$ such that $\mathcal{Q} \neq \mathcal{Q}'$ and $\mathcal{A}^{\text{ext}}_{\mathcal{Q}'} = \mathcal{A}^{\text{ext}}_{\mathcal{Q}}$. Since $\mathcal{Q} \neq \mathcal{Q}'$ it implies either of the two cases

- **Case 1:** There exists a $q' \in \text{ext}(\mathcal{Q}')$ such that $q' \notin \text{ext}(\mathcal{Q})$. This implies that

$$\exists\, a^*_{q'} \in \mathcal{A}^{\text{ext}}_{\mathcal{Q}'} \quad \text{and} \quad \exists\, a^*_{q'} \in \mathcal{A}^{\text{ext}}_{\mathcal{Q}} \qquad (a^*_{q'} \text{ is unique for all } q' \in \Delta(\mathcal{O}))$$

  This results contradicts $\mathcal{A}^{\text{ext}}_{\mathcal{Q}'} = \mathcal{A}^{\text{ext}}_{\mathcal{Q}}$.

- **Case 2:** There exists a $q \in \text{ext}(\mathcal{Q})$ such that $q \notin \text{ext}(\mathcal{Q}')$. We follow the same reasoning as Case 1, i.e.,

$$\exists\, a^*_{q} \in \mathcal{A}^{\text{ext}}_{\mathcal{Q}} \quad \text{and} \quad \exists\, a^*_{q} \in \mathcal{A}^{\text{ext}}_{\mathcal{Q}'} \qquad (a^*_{q} \text{ is unique for all } q \in \Delta(\mathcal{O}))$$

  resulting in a contradiction with $\mathcal{A}^{\text{ext}}_{\mathcal{Q}'} = \mathcal{A}^{\text{ext}}_{\mathcal{Q}}$.

Hence $\mathcal{A}^{\text{ext}}$ is a unique representation for all credal sets. □

D.2 PROOF OF THEOREM 5.4

Proof. We know that for s_θ to be strictly proper, the following must hold for all beliefs $\mathcal{P} \subseteq \Delta(\mathcal{O})$

$$V^{\mathcal{P}}_\theta(\mathcal{P}) = V^{\mathcal{P}}_\theta(\mathcal{Q}) \quad \text{iif} \quad \mathcal{P} \simeq \mathcal{Q}$$

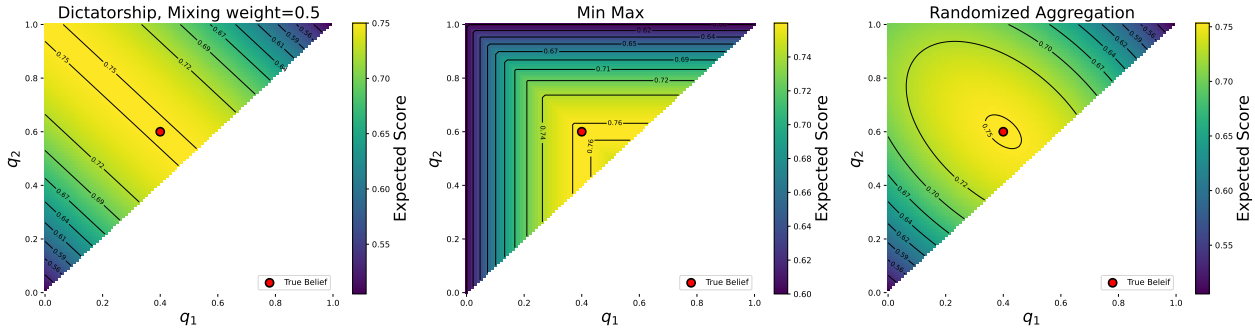
($\Rightarrow$) Given, $\theta \in \Delta(\boldsymbol{\rho})$ has full support, $V^{\mathcal{P}}_\theta(\mathcal{P}) = V^{\mathcal{P}}_\theta(\mathcal{Q})$ implies that,

$$\begin{aligned}
\rho(\{\mathbb{E}_p[s_\rho(\mathcal{P}, o)]\}_{p\in\mathcal{P}}) &= \rho(\{\mathbb{E}_p[s_\rho(\mathcal{Q}, o)]\}_{p\in\mathcal{P}}) \quad \forall \rho \in \boldsymbol{\rho} \quad \forall \mathcal{P} \subseteq \Delta(\mathcal{O}) \\
\boldsymbol{\lambda}^\top\{\mathbb{E}_p[u(a^*_{\boldsymbol{\lambda}^\top\mathcal{P}}, o)]\} &= \boldsymbol{\lambda}^\top\{\mathbb{E}_p[u(a^*_{\boldsymbol{\lambda}^\top\mathcal{Q}}, o)]\} \quad \forall \boldsymbol{\lambda} \in \Delta^{|ext(\mathcal{P})|} \quad \forall \mathcal{P} \subseteq \Delta(\mathcal{O}) && (\boldsymbol{\rho}\text{: fixed linear aggregation}) \\
\boldsymbol{\lambda}^\top\{\mathbb{E}_p[u(a^*_{\boldsymbol{\lambda}^\top\mathcal{P}}, o)]\} &= \boldsymbol{\lambda}^\top\{\mathbb{E}_p[u(a^*_{\boldsymbol{\lambda}^\top\mathcal{Q}}, o)]\} \quad \forall \boldsymbol{\lambda} \in \{\delta_i\}_{i\in|ext(\mathcal{P})|} \quad \forall \mathcal{P} \subseteq \Delta(\mathcal{O}) && (\{\delta_i\}_{i\in|ext(\mathcal{P})|} \subset \Delta^{|ext(\mathcal{P})|}) \\
\mathbb{E}_p[u(a^*_p, o)] &= \mathbb{E}_p[u(a^*_q, o)] \quad \forall p \in \mathcal{P} \quad \forall \mathcal{P} \subseteq \Delta(\mathcal{O}) && (q := \delta_i^T\mathcal{Q}) \\
a^*_p &= a^*_q \quad \forall p \in \mathcal{P} \quad \forall \mathcal{P} \subseteq \Delta(\mathcal{O}) \\
& && \text{(Using Lemma 4.10 as } s_\theta \text{ is strictly proper for precise distributions)} \\
\implies \mathcal{A}^{ext}_{\mathcal{P}} &= \mathcal{A}^{ext}_{\mathcal{Q}} && (\text{By Definition of } \mathcal{A}^{ext}) \\
\implies \mathcal{P} &\simeq \mathcal{Q} && (\text{Lemma 5.3})
\end{aligned}$$

($\Leftarrow$) Given that $\mathcal{P} \simeq \mathcal{Q}$ we show that $V^{\mathcal{P}}_\theta(\mathcal{P}) = V^{\mathcal{P}}_\theta(\mathcal{Q})$. This is trivial since two equivalent forecasts produce the same underlying partial order on the actions $\mathcal{A}$. As aggregation functions make this partial order complete, by the property of being a function, they will result in the same complete order for the same partial order. Therefore, given $\mathcal{P} \simeq \mathcal{Q}$ implies that

$$\begin{aligned}
V^{\mathcal{P}}_\rho(\mathcal{P}) &= V^{\mathcal{P}}_\rho(\mathcal{Q}) \quad \forall \rho \in \boldsymbol{\rho} \\
\mathbb{E}_\theta[V^{\mathcal{P}}_\rho(\mathcal{P})] &= \mathbb{E}_\theta[V^{\mathcal{P}}_\rho(\mathcal{Q})] \quad \forall \theta \in \Delta(\boldsymbol{\rho}) \\
V^{\mathcal{P}}_\theta(\mathcal{P}) &= V^{\mathcal{P}}_\theta(\mathcal{Q}) \quad \forall \theta \in \Delta(\boldsymbol{\rho})
\end{aligned}$$

Therefore, the imprecise forecaster is truthful in the epistemic sense w.r.t the strictly proper IP scoring rule s_θ. □



(a) (b) (c)

Figure 3: (Left-to-right) In the first figure we simulate the scoring rule where ρ is a dictatorship with a fixed mixing weight of 0.5, in the middle figure we simulate the scoring rule with min-max ρ (pessimistic decision maker) and in the last figure we simulate the scoring rule where the aggregation is a randomized dictatorship and the forecaster obtains a distribution $\theta = \mathcal{U}[0,1]$ over ρ. The lower half of the figure is not plotted since that corresponds to region $q_1 > q_2$, i.e. lower probability being greater than upper probability

E SIMULATIONS

To test the sanity of our proposed scoring rule, we simulate a scenario where an imprecise forecaster predicts a binary outcome (e.g., chance of rain tomorrow). We assume the forecaster has an imprecise forecast $[0.4, 0.6]$ and uses an imprecise scoring rule s_ρ where ρ is a dictatorship or some other aggregation like min-max. We compare this to our randomized imprecise scoring rule s_θ. Given the binary outcome, the forecaster reports an interval $\mathcal{Q} := [q_1, q_2]$ where q_1 denotes the lower probability and q_2 the upper probability respectively. Figure 3 highlights that the randomized scoring rule s_θ is strictly proper for imprecise forecasts as it has the highest expected score for the forecaster only when the forecaster reports his true belief. While in other cases of using a deterministic imprecise scoring rule s_ρ, if DM provides a ρ such that it is a dictatorship, such as in the case of Figure 3(a), the scoring rule is proper; however, the forecaster can lie by reporting the dictator. This can be inferred from the contour that the point $[0.5, 0.5]$, which corresponds to the precise forecast 0.5, also has the highest expected score. With ρ being a min-max rule, the scoring rule s_ρ is proper but not strictly as other imprecise forecasts allow the forecaster to obtain the same expected score. For our implementation we consider $\mathcal{A} = [0,1]$ and $u(a,o) := (o-a)^2$ to satisfy Lemma 4.10.