Journal of Machine Learning Research 27 (2026) 1-52 Submitted 5/25; Published 3/26

Three Types of Calibration using Properties and their Semantic and Formal Relationships

Rabanus Derr RABANUS.DERR@UNI-TUEBINGEN.DE
University of Tübingen and
Tübingen AI Center
Tübingen, 72076, Germany

Jessie Finocchiaro FINOCCH@BC.EDU
Computer Science,
Boston College
Chestnut Hill, MA 02467, USA

Robert C. Williamson BOB.WILLIAMSON@UNI-TUEBINGEN.DE
University of Tübingen and
Tübingen AI Center
Tübingen, 72076, Germany

Editor: Mehryar Mohri

Abstract

Fueled by discussions around "trustworthiness" and algorithmic fairness, calibration of predictive systems has regained scholars' attention. The vanilla definition and understanding of calibration is, simply put, on all days on which the rain probability has been predicted to be p, the actual frequency of rain days was p. However, the increased attention has led to an immense variety of new notions of "calibration." Some of the notions are incomparable, serve different purposes, or imply each other. In this work, we provide two accounts which motivate calibration: self-realization of forecasted properties and precise estimation of incurred losses of the decision makers relying on forecasts. We substantiate the former via the reflection principle and the latter by actuarial fairness. For both accounts we formulate prototypical definitions via properties Γ of outcome distributions, e.g., the mean or median. The prototypical definition for self-realization, which we call Γ-*calibration*, is equivalent to a certain type of swap regret under certain conditions. These implications are strongly connected to the omniprediction learning paradigm. The prototypical definition for precise loss estimation is a modification of *decision calibration* adopted from Zhao et al. (2021). For binary outcome sets both prototypical definitions coincide under appropriate choices of reference properties. For higher-dimensional outcome sets, both prototypical definitions can be subsumed by a natural extension of the binary definition, called *distribution calibration* with respect to a property. We conclude by commenting on the role of groupings in both accounts of calibration often used to obtain multicalibration. In sum, this work provides a semantic map of calibration in order to navigate a fragmented terrain of notions and definitions.

Keywords: calibration, property elicitation, property identification, Bayes risk, swap regret

1. Introduction

Calibration has increasingly gained interest since it seems to provide a mathematical criterion of "trustworthy"[1] predictions (Noarov and Roth, 2024) and it is a major component of studies on algorithmic fairness (Chouldechova, 2017; Hébert-Johnson et al., 2018). Furthermore, the advent of capable, deep learning techniques gave rise to investigations of calibration of general deep neural networks (Guo et al., 2017) and large language models (Cruz et al., 2024; OpenAI (2023), 2024; Kalai and Vempala, 2024).

While calibration has become a quantity of interest in empirical studies (OpenAI (2023), 2024; Perdomo et al., 2023; Cruz et al., 2024), theoretical works came up with dozens of new (and old) definitions of calibration investigating particular relationships between them. To name a few: Γ-calibration (Noarov and Roth, 2023), decision calibration (Zhao et al., 2021), U-calibration (Kleinberg et al., 2023), class-wise calibration (Kull et al., 2019), confidence calibration (Guo et al., 2017), Global Interpretable Calibration Index (Cabitza et al., 2022), distance to calibration (Błasiok et al., 2023), smooth calibration (Foster and Hart, 2018). The "jungle" of current notions of calibration is difficult to navigate.

As a result, wide-spread use of the term "calibration" has blurred the boundaries of what is meant by it. However, all usages have in common that calibration is either a criterion to judge predictions in light of obtained data or the process of fulfilling such a criterion. In this work we stick to the former (cf. (Höltgen and Williamson, 2023)). Our goal is *not* to provide an exhaustive list of notions of calibration.[2] Instead, we follow our main question: **What is the abstract purpose of calibration?** This way we provide a semantic map and guide through key notions of calibration and their central relationships.

Calibration is often contrasted with pure predictive accuracy (Cruz et al., 2024; Van Calster et al., 2019; Seidenfeld, 1985). While predictive accuracy guarantees the "usefulness" of predictions, calibration guarantees the "trustworthiness" of the predictions. The "usefulness" narrative is readily justified. Expected risk minimization, the core principle behind a large portion of learning techniques, is the negative analogue of expected utility maximization. The lower the risk, the more useful the predictions are, when measured with the corresponding loss (respectively negative utility) function.

The "trustworthiness"-narrative, however, requires a more detailed explanation. We identified two accounts of calibration which could justify it:

Self-realization: For the instances where some value c was forecasted, the actual outcomes can be summarized to a value close to c.

Precise Loss Estimation: The forecasted values let one provide estimates of incurred losses (for certain loss functions) which are close to the actual materialized losses.

While these accounts are not an exhaustive list, these two paradigms account for the majority of calibration motivations. We found a third account of calibration in the literature which

1. We hesitate to attempt a concise definition of trustworthiness in this work for the ambiguity and vagueness of its purpose and meaning. Instead, we write "trustworthiness" to emphasize to the reader that the term itself is and should be loaded with more than the formal intuitions for some facets of trustworthiness presented in this work.
2. After trying this for some time, we gave up on this journey due to the immense variety and subtleties of the suggested variants of calibration.

focuses on the approximate equivalence of means. Here, predictions are calibrated if the average of outcomes is equal to the average of predictions on certain subgroups or even individuals (Dawid, 1985; Zhao et al., 2020; Luo et al., 2022; Höltgen and Williamson, 2023). This account is somewhat located between the narratives of "usefulness" and "trustworthiness." Note that all accounts extend on a certain facet of the binary vanilla calibration definition (Definition 1) which has historically motivated previous definitions.

To uncover the two central accounts of this work, we rewrite current definitions of calibration using the language of properties (Osband, 1985; Fissler and Ziegel, 2016; Lambert et al., 2008). Properties, simply put, are functions from distributions on the outcome set $\mathcal{Y}$ to some value set $\mathcal{R}$, e.g., $\mathcal{R} = \mathbb{R}$ or $\mathcal{R} = \Delta(\mathcal{Y})$. We distinguish between *optimization-level* and *decision-level* properties. Optimization-level properties specify the actual predicted entity, e.g., full distribution, mean, α-quantile, or best action. Decision-level properties most often correspond to downstream uses of the optimization-level property, such as decision makers informing their discrete action via a prediction about outcome probabilities, or ranking of classes being used to deduce the top-k most likely outcomes. Decision-level properties borrow their name since they generalize maximum expected utility decision makers. Notably, decisions are often discrete, and therefore hard to optimize directly, necessitating the distinction between (often continuous) optimization-level properties and (often discrete) decision-level properties. Properties subsume not only utility-based decision makers, but rather arbitrary statistical properties of distributions as well, such as quantiles[3], ratios of expectations, or class marginals, to name a few. Furthermore, the elicitability and identifiability of properties as detailed in Section 3 relate properties to loss functions and their optimization criterion. This makes properties a useful vehicle to study calibration in the abstract.

Our contributions are summarized as follows:

1. We subsume many existing definitions under three core *types* of definitions of calibration: distribution calibration with respect to a (decision-level) property Φ (Definition 2), property calibration with respect to an abstract property Γ (Definition 5) and decision calibration with respect to a set of loss functions $\mathcal{L}$ (Definition 9).

2. We show that all types of definitions of calibration collapse in the case of binary outcome sets and appropriate choices of Φ, Γ and $\mathcal{L}$ (Proposition 11). More generally, we argue that distribution calibration is the central "parent" notion which implies both of the others (Proposition 6 and Proposition 12), as summarized in Figure 1. When generalizing to the approximate case, see Propositions 23 and 31 and Figure 6 in the appendix.

3. We elaborate on property calibration as a prototypical definition for the account of self-realization. We show it can equivalently be expressed as a type of swap regret (Proposition 7 and Proposition 26). Furthermore, we prove that self-realization is inherited by *refined* properties— those properties Φ that can be defined by composing a function with some original property Γ (Proposition 8 and Proposition 28), a characteristic of the notion which is exploited in omniprediction (cf. Proposition 30).

3. Calibrated quantiles are relevant for conformal predictions (Jung et al., 2023).

4. We contextualize decision calibration as a prototypical notion of precise loss estimation. In particular, we show it requires simultaneous precise Bayes risk estimation for several loss functions of interest (Proposition 15). Furthermore, we dissect the relationship between self-realization and precise loss estimation, concluding that these accounts are incommensurable (Section 6.3).

5. We provide a non-exhaustive categorization of existing notions of calibration in the three types in Table 1 and Table 2.

The distinction of the three types of definitions of calibration and their relationship to the presented accounts of self-realization and precise loss estimation has, to the best of our knowledge, not been made in the literature. Properties have been used to formalize calibration in (Gneiting and Resin, 2023) and (Noarov and Roth, 2023). Some of the formal relationships have been discussed already in literature (e.g., distribution calibration implies decision calibration has been argued for in (Zhao et al., 2021)), but not within the more general picture of properties.

In this work, we pay attention on how to define calibration beyond binary outcomes. We largely ignore another strand of work on calibration notions which focus on how to approximate (binary) Vanilla calibration. Different choices (e.g., "distance to calibration" (Błasiok et al., 2023), "calibration decision loss" (Hu and Wu, 2024), "cutoff calibration error" (Rossellini et al., 2025), "smooth calibration" (Foster and Hart, 2018), "projected smooth calibration" (Gopalan et al., 2024a)) yield different decision-theoretic and empirical approximability guarantees. Rossellini et al. (2025); Haghtalab et al. (2024) give a more in-depth, axiomatic discussion about calibration error metrics. Our treatment is largely agnostic to the choice of metrics to capture the approximation, though the relationship between properties and the existence of calibration metrics satisfying normative axioms is an interesting line of future work.

Figure 1 graphically summarizes the two strands which we will follow in this work. The figure depicts the idealized "perfect" calibration case. For approximate notions the implications require additional assumptions (Figure 6). Clearly, we don't expect the reader to yet make sense of the definitions and implications.

2. An Exemplary Motivation

Let us walk through an example to get an intuitive understanding of the subtleties of different formalizations of calibration.

Perdomo et al. (2023) study the Dropout Early Warning System (DEWS) implemented at Wisconsin Public Schools. Based on state level testing in the 8th grade, the system predicted on-time high school graduation likelihood. Hence, the DEWS predictions themselves, i.e., the *optimization-level property*, are probability forecasts on a binary outcome set. As the authors observe, the forecasts are not perfectly calibrated following *binary vanilla calibration* (Definition 1), i.e., the graduation rate of students who get a certain prediction p is not p.

The predictions of the DEWS are given to the department authorities who define three risk categories "high risk," "medium risk," and "low risk" for dropout. The risk category determines the measures to be taken to increase the graduation likelihood. In other words, the risk categories form the *decision-level property*. Are the *predictions calibrated conditioned on*

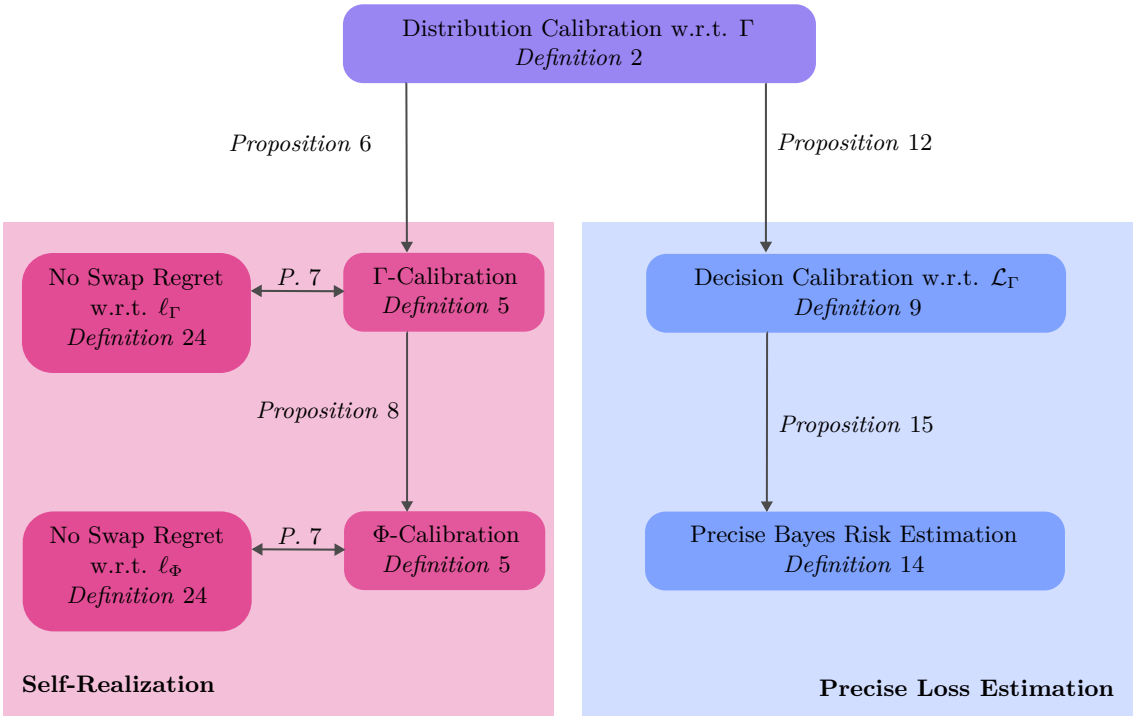


Figure 1: Relationships between Notions of Calibration. Implications under perfect calibration, finite $\mathcal{Y}$ and elicitable property Γ and Φ. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

the risk categories, i.e., is the average graduation rate among all persons who are categorized as "high risk" (or "medium risk" or "low risk") equal to the average prediction score of those persons? This property, which we call *distribution calibration with respect to the risk categories* (Definition 2), is not fulfilled as immediately derived from (Perdomo et al., 2023, Figures 1 and 2).

But, the average graduation rate among all persons who are categorized as "high risk" is lower than the average graduation rate in all other categories (for the other categories respectively). In other words, the "high risk" category self-realizes. Hence, the predictions are *property calibrated* (Definition 5).

Finally, self-realization of the decision-level property is not necessarily the intention of calibration. Instead one can ask whether the department authorities can estimate their expected loss in mis-assigning students to the wrong group. For this one would need to introduce a loss function, whose minimizer is the risk category assignment. Note that several such loss functions exist. Depending on the choice of the loss function, the predictions are precisely estimating the loss, i.e., the predictions are *decision calibrated* (Definition 9), or not.

Exactly, how calibrated the DEWS predictions are is irrelevant for the lessons we want to convey in this section. Some of the notions of calibration are fulfilled to a lesser degree than others. We use this example to demonstrate that there is no "right" definition of calibration. This leaves a choice open which type and definition of calibration to consider in

a contextualized problem. This choice has semantics and implications which we disentangle in the following.

3. Formal Setup

Data Distribution Let $(\Omega, \Sigma, \lambda)$ be a standard probability space. Let $\mathcal{X}$ (and $\mathcal{Y}$) be an input set (and outcome set) respectively. For the sake of simplicity, we assume that those sets are finite dimensional Euclidean spaces or finite sets, equipped with the standard Borel-σ-algebra $\mathcal{B}(\mathcal{X})$ (respectively $\mathcal{B}(\mathcal{Y})$). We define two random variables $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$. The distribution induced on $\mathcal{X} \times \mathcal{Y}$ by the joint random variable (X, Y) is denoted D, and called the *data distribution*. The set of all data distributions is denoted $\Delta(\mathcal{X} \times \mathcal{Y})$. For the expected value of a measurable function $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ we write $\mathbb{E}_{(X,Y)\sim D}[g(X, Y)]$ or simply $\mathbb{E}_D[g(X, Y)]$.

The marginals are denoted D_X (respectively D_Y). We write push-forward measures such as for a measurable map $f\colon \mathcal{X} \to \mathbb{R}$ as $D_{f(X)}$. Conditional distributions, e.g., conditional probability of Y given X are written as $D_{Y|X\in A}$, where $A \in \mathcal{B}\mathcal{X}$ and $D_X(A) > 0$. For the corresponding conditional expectation for a measurable function $g\colon \mathcal{Y} \to \mathbb{R}$ we write,

$$\mathbb{E}_D\left[g(Y)|X \in A\right] \coloneqq \mathbb{E}_{Y\sim D_{Y|X\in A}}\left[g(Y)\right].$$

We pay attention to not conditioning on measure 0 events. For that reason, we refer to the support of a measure, $\operatorname{supp} D_{f(X)}$, which is the set $\{r \in \mathbb{R}\colon D_{f(X)}(\{r\}) > 0\}$. By $\mathcal{P}$ we mean a subset of $\Delta(\mathcal{Y})$ the set of all probability distributions on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. We call a data distribution D *regular with respect to* $\mathcal{P}$ if and only if for every $A \in \mathcal{B}(\mathcal{X})$ it is true that $D_{Y|X\in A} \in \mathcal{P}$. Finally, we denote the Iverson-brackets as $[\![\cdot]\!]$, i.e., $[\![A]\!] = 1$ if A is true, $[\![A]\!] = 0$ otherwise.

Properties Let $\mathcal{R}$ be a set of property values equipped with a metric m; i.e., $(\mathcal{R}, m)$ forms a metric space. Generally, we see three common choices of value sets: (i) if $\mathcal{R} \subseteq \mathbb{R}$ is an interval, we set m to be the absolute difference, (ii) if $|\mathcal{R}| < \infty$, we set m to be the discrete metric. Finally, (iii) if $\mathcal{R} = \Delta(\mathcal{Y})$, we set m to be the total variation distance. Furthermore, $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$ is a measurable space for $\mathcal{B}(\mathcal{R})$ being the Borel-σ-algebra induced by the topology given via m. A measurable function $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called a *property*.[4] Without loss of generality we assume that Γ is surjective. To distinguish between optimization-level and decision-level properties we sometimes use Γ and Φ respectively. In Section 6, we additionally use Θ to denote Bayes risk properties (defined later, Definition 13).

A property is *continuous* iff Γ is continuous with respect to the total variation distance on $\mathcal{P}$ and the metric m on $\mathcal{R}$. In particular, a property is *Lipschitz continuous* iff Γ is Lipschitz continuous with respect to the total variation distance on $\mathcal{P}$ and the metric m on $\mathcal{R}$.

A property $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called *elicitable* if there exists a loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ which is measurable in the first variable for all fixed $\gamma \in \mathcal{R}$ in the second variable and such that,

$$\Gamma(P) \in \arg\min_{\gamma\in\mathcal{R}} \mathbb{E}_{Y\sim P}[\ell(Y, \gamma)],$$

4. Note that measurability can be guaranteed by inducing a σ-algebra on $\mathcal{P}$ or simply referring to the Borel-σ-algebra on $\Delta(\mathcal{Y})$ induced by the total variation distance, in particular if $\mathcal{Y}$ is finite.

for all $P \in \mathcal{P}$. In the case ℓ elicits Γ, we say the loss ℓ is $\mathcal{P}$-consistent with respect to Γ. Overloading notation, we sometimes write $\ell(P, \gamma) = \mathbb{E}_{Y\sim P}[\ell(Y, \gamma)]$. An elicitable property can be understood as a "best response" for a minimum expected loss decision maker.

A property $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called *identifiable* if there exists an identification function $V\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ which is measurable in both variables, and

$$\Gamma(P) = \gamma \Leftrightarrow \mathbb{E}_{Y\sim P}[V(Y, \gamma)] = 0,$$

for all $P \in \mathcal{P}$. We overload notation and write $V(P, \gamma) = \mathbb{E}_{Y\sim P}[V(Y, \gamma)]$.

A measurable function $f\colon \mathcal{X} \to \mathcal{R}$ is called a Γ-*predictor* for the property $\Gamma : \mathcal{P} \to \mathcal{R}$. Properties, as well as property predictors can be simply stacked to vectors, e.g., the distribution predictor $f\colon \mathcal{X} \to \Delta(\mathcal{Y})$ is a $(\Gamma_y)_{y\in\mathcal{Y}}$-predictor, where Γ_y denotes the property: probability of outcome $y \in \mathcal{Y}$.

4. Distribution Calibration

Calibration has been well-studied long before its evolution as a "trustworthiness" criterion or "fairness" criterion (as multi-calibration) in machine learning (Murphy and Epstein, 1967; Murphy and Winkler, 1977; Dawid, 1982; DeGroot and Fienberg, 1983).[5][6] Calibration was considered part of a canon of "goodness"-measures of (meteorological) forecasts (Murphy and Epstein, 1967; DeGroot and Fienberg, 1983). In particular, calibration captured the following intuition: if it rains a p-proportion of times on the days on which the prediction is p for rain, the predictions are calibrated. Definition 1 makes this formal.

Definition 1 (Binary Vanilla Calibration)**.** *Let $\mathcal{X} \subseteq \mathbb{R}$ be an input set and $\mathcal{Y} = \{0, 1\}$ a binary outcome set. A predictor $f\colon \mathcal{X} \to [0, 1]$ predicting the mean, i.e., the probability $D_{Y|X=x}(Y = 1) = \mathbb{E}[Y|X = x]$, is* calibrated *on a distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$[7],*

$$\mathbb{E}_{(X,Y)\sim D}[Y|f(X) = \gamma] = \gamma.$$

Consider $f\colon \mathcal{X} \to \mathcal{P}$ to be a full distribution forecast for a finite $\mathcal{Y}$. It is not always relevant, nor attainable to provide full distribution estimates which are calibrated conditioned on the distributional predictions. In particular, in high-dimensional class probability prediction problems it is not realistic to achieve reasonable calibration guarantees (Roth and Shi, 2024; Noarov and Roth, 2024). The number of events to condition on grows exponential in the dimensionality of the outcome set $\mathcal{Y}$. Hence, scholarship suggested conditioning on "decision-events" (Noarov et al., 2023; Roth and Shi, 2024), marginals (Kull et al., 2019), or top-labels (Guo et al., 2017; Gupta and Ramdas, 2021). In other words, the number of conditioning events is reduced by focusing only on the "relevant" ones.[8]

5. In some of the older literature the terms "reliability" and "validity" were (inconsistently) used instead of "calibration".
6. Murphy and Winkler (1977) provide the first calibration plots, objective frequency versus prediction.
7. We refer to the support of the push-forward measure to exclude that we condition on measure zero sets. In other words, our calibration conditions focus only on substantially often predicted values. That has the side-effect that a predictor which predicts each value only on measure zero sets is directly calibrated. Note that this is a theoretical artifact which does not play a role for any empirical distribution D.
8. An interesting perspective on the conditioning events is given in (Grünwald, 2018, Section 2.3). The author essentially describes calibrated predictors as "safe" summaries of a certain set of probability distributions,

We subsume all those notions via a specified property. For the sake of readable notation we use Γ as the symbol for the property here, even though it can refer to a decision-level or optimization-level property. For instance, Γ could be the map which maps all distributions to one of their marginals (cf. (Kull et al., 2019)) or best responses with respect to some utility function (cf. (Noarov et al., 2023)).

Definition 2 (Distribution Calibration with Respect to Γ)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, where $\mathcal{Y}$ is finite. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. The predictor f is* distribution calibrated with respect to Γ *on D if for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,*[9]

$$\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = \gamma] = \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = \gamma], \quad \forall y \in \mathcal{Y},$$

where $f_y(x) \in [0,1]$ denotes the y-component of the prediction $f(x) \in \Delta(\mathcal{Y})$ for $x \in \mathcal{X}$. The predictor f is $\alpha(\gamma)$-approximate distribution calibrated with respect to Γ *on D if for every $\gamma \in \operatorname{im}\Gamma \circ f$, $\alpha(\gamma) \in \mathbb{R}$,*

$$|\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \Gamma \circ f(X) = \gamma]| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y}.$$

We illustrate Definition 2 in Figure 2, where we demonstrate exactly (left) and approximately (right) satisfying distribution calibration with respect to the property $\Gamma(p) = \arg\max_y p_y$ representing the mode.

Note that, in principle, $\mathcal{Y}$ does not need to be finite. However, it simplifies notation and facilitates understanding. We provide definitions and results for general $\mathcal{Y}$ in Appendix A. Note that if $\mathcal{Y}$ is infinite, most predictions actually directly refer to properties and not the full distribution (e.g., in Gaussian Process Regression, the mean and covariance are estimated, which are properties of the full distribution).

4.1 Distribution Calibration is Inherited

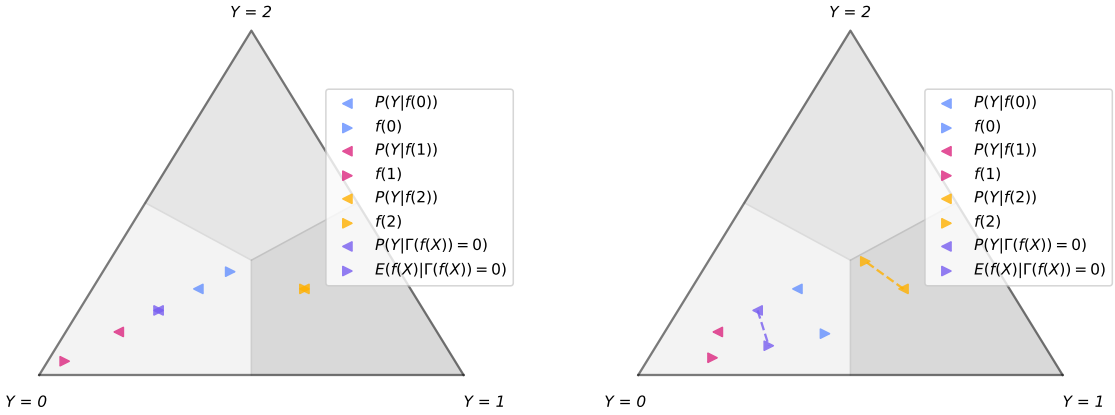
Distribution calibration with respect to some property is naturally inherited by *refined properties*. A property Φ is refined by another property Γ, if there exists a mapping which applied on Γ gives Φ.

Definition 3 (Property Refinement (Frongillo and Kash, 2021, Definition 12))**.** *A property $\Phi\colon \mathcal{P} \to \mathcal{R}'$ is called* refined by Γ*, if $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is a property such that there exists a function $\phi\colon \mathcal{R} \to \mathcal{R}'$ with $\Phi = \phi \circ \Gamma$.*[10]

Proposition 4 (Distribution Calibration is Inherited)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, where $\mathcal{Y}$ is finite. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor . Suppose, furthermore, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$.*[11]

in our case the observed distribution D. The choice of conditioning events is then tantamount to a choice of uses for which the predictor is a "safe probability".

9. Since $\mathcal{Y}$ is finite, we have $\mathcal{P} \subseteq \mathbb{R}^{|\mathcal{Y}|}$. Hence, the distribution calibration condition is equivalent to $D_{Y|\Gamma \circ f(X) = \gamma} = \mathbb{E}_D[f(X) | \Gamma \circ f(X) = \gamma]$.
10. Property refinement is intricately related to the notion of *indirect property elicitation* (Frongillo and Kash, 2021; Finocchiaro et al., 2024, 2021).
11. This technical assumption is necessary to exclude problems with aggregations of measure zero sets related to our definitional choice that our calibration conditions only hold almost everywhere. The assumption is essentially equivalent to assuming that f maps to at most countably many different values. The assumption re-appears in Proposition 8 and Proposition B.3.



(a) Perfect distribution calibration with respect to Γ.

(b) Approximate distribution calibration with respect to Γ.

Figure 2: Illustration of distribution calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $\mathcal{X} = \{0, 1, 2\}$. We define $\Gamma(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ are drawn in different shades of gray. The left-directing markers denote the true conditional distribution for different choices of $x \in \mathcal{X}$. The right-directing markers denote the predicted distribution by a predictor f. The purple markers are convex combinations of the blue and red markers, where the convex combination is defined through the marginal distribution on $\mathcal{X}$ which is in our case fixed to be uniform. The dashed lines highlight the deviation from the true outcome distribution conditioned on a value of $\Gamma \circ f$ versus the expected forecast conditioned on a value of $\Gamma \circ f$. Only the forecasts change when comparing Figure 2a versus Figure 2b.

If the predictor f is $\alpha(\gamma)$-approximate distribution calibrated with respect Γ on D, then it is $\alpha'(c)$-approximate distribution calibrated with respect to $\Phi \coloneqq \phi \circ \Gamma$ for all $\phi\colon \mathcal{R} \to \mathcal{R}'$ where $\alpha'(c) \coloneqq \sup_{\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}\colon \phi(\gamma)=c} \alpha(\gamma)$.

Proof By assumption, for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$|\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \Gamma \circ f(X) = \gamma]| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y}.$$

Hence, in particular, for all $c \in \operatorname{supp} D_{\Phi \circ f(X)}$ and all $y \in \mathcal{Y}$,

$$\begin{aligned}
\alpha'(c) &= \sup_{\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}\colon \phi(\gamma)=c} \alpha(\gamma) \\
&\geq \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [\alpha(G) \mid \phi(G) = c] \\
&= \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [|\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = G] - \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = G]| \mid \phi(G) = c] \\
&\geq \left| \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = G] - \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = G] \mid \phi(G) = c] \right| \\
&\geq |\mathbb{E}_D[[\![Y = y]\!] | \Phi \circ f(X) = c] - \mathbb{E}_D[f_y(X) | \Phi \circ f(X) = c]| \,.
\end{aligned}$$

Note that the supremum in the first line is well-defined, because for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. ■

Under mild geometric conditions on the predictions, perfect distribution calibration with respect to all elicitable binary properties implies distribution calibration (Appendix C). This statement can be understood as a reverse implication to Proposition 4. Figure 3 illustrates the idea of property refinement and inheritance of distribution calibration.

Distribution calibration with respect to Γ forms the entry point to the two accounts of calibration mentioned earlier (cf. Figure 1 and Figure 6). Distribution calibration is close to the account of mean-matching briefly mentioned in the introduction. However, this account puts more focus on groups and individuals, i.e., information provided through the input $\mathcal{X}$ (cf. (Höltgen and Williamson, 2023)).

5. Γ-Calibration as Self-Realization of Predictions

Central to vanilla calibration is the self-referential aspect in the conditioning, i.e., given the predicted Bernoulli distribution has parameter γ, the actual distribution of outcomes is Bernoulli-distributed with parameter γ. In this section we generalize calibration around this perspective.[12]

As observed by Jung et al. (2023, 2021) and Gupta et al. (2022), the expectation operator in Definition 1 can be replaced by moments, respectively quantile functions. From these works, Noarov and Roth (2023) distilled a general definition of calibration with respect to properties. Different to Noarov and Roth (2023) we ignore groupings based on input X. Hence, we consider Γ-calibration and not "multi"-Γ-calibration.

12. The self-referential aspect is similar in nature to a widely discussed principle for deference of belief in philosophy, the *reflection principle* (Molinari, 2023). Van Fraassen (1984) introduced the reflection principle to argue that reflection of subjective beliefs of future selves is a requirement for rationality of the agent. In a footnote, Seidenfeld (1985, p. 276) suggest a definition of "subjective calibration" which resembles our Definition 5 and the reflection principle as defined in (Van Fraassen, 1984). Unfortunately, we were not able to identify the work the author refers to.

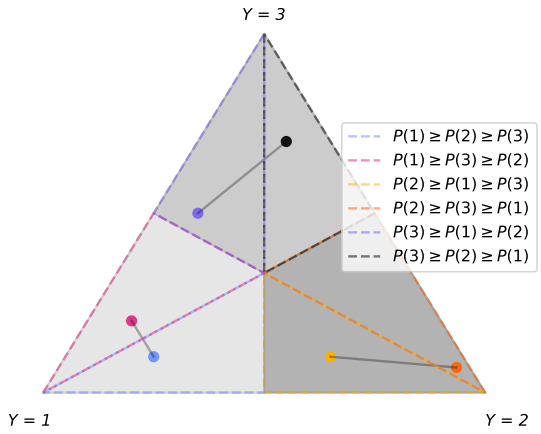


Figure 3: Illustration of property refinement and inheritance of distribution calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $|\mathcal{X}| = 6$. We define $\Gamma(P) = (y_1, y_2, y_3)$ such that $P(Y = y_1) \geq P(Y = y_2) \geq P(Y = y_3)$ and $\Phi(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ have colored boundaries listed in the legend. The level sets of Φ are drawn in different shades of gray. The property Γ refines Φ. We assume that f is a distributional predictor whose predictions are marked as dots. The color of the dots represents $\Gamma \circ f(x)$. The lines indicate the convex combination of predictions happening when conditioning on $\Phi \circ f(X)$ instead of $\Gamma \circ f(X)$. Since the level sets of Φ are all convex the line is always contained *within* a single level set.

Definition 5 (Γ-Calibration (Noarov and Roth, 2023))**.** *Suppose* $(\mathcal{R}, m)$ *is a metric space. Let* $\Gamma \colon \mathcal{P} \to \mathcal{R}$ *be a property and* D *a data distribution on* $\mathcal{X} \times \mathcal{Y}$ *regular wrt.* $\mathcal{P}$*. Let* $f \colon \mathcal{X} \to \mathcal{R}$ *be a* Γ*-predictor. Then,* f *is* Γ-calibrated *on* D *if for every* $\gamma \in \operatorname{supp} D_{f(X)}$*,*

$$\Gamma(D_{Y|f(X)=\gamma}) = \gamma.$$

The Γ*-predictor* f *is* $\alpha(\gamma)$-approximate Γ-calibrated *on* D *if for every* $\gamma \in \operatorname{supp} D_{f(X)}$*,*

$$m(\Gamma(D_{Y|f(X)=\gamma}), \gamma) \leq \alpha(\gamma).$$

Central to the definition is the self-realization aspect. The property of the distribution of outcomes, on which the predictor f predicted γ, actually is γ. The reader familiar with (Noarov and Roth, 2023) might question the use of a general property Γ in Definition 5. For a discussion see Appendix D.

Gneiting and Resin (2023) introduced "T-calibration" (Definition 2.7 therein), which is a measure-theoretic definition of Γ-calibration. The authors already note that certain properties Γ are not sensible for calibration following (Noarov and Roth, 2023, Definition 3.2), but only give necessary not sufficient conditions for the sensibility (cf. Appendix D).

In practice it is rarely the case that a Γ-predictor is perfectly Γ-calibrated. However, there are algorithms, such as (Noarov and Roth, 2023, Algorithm 1) which post-hoc approximately

calibrate Γ-predictors (for identifiable Γ). In that context, remember our conventions for m listed in Section 3, e.g., if $\mathcal{R}$ is an interval on the real line, then m is set to be the absolute difference for simplicity. However, the choice of metric m is worthy of its own line of inquiry (cf. Section 1).

In addition, observe that in many existing definitions the authors commit to a specified aggregation of $\alpha(\gamma)$ over all $\gamma \in \mathrm{im} f$. For instance, (Noarov and Roth, 2023) consider the expected squared value, $\alpha \coloneqq \mathbb{E}_{\gamma \sim D_{f(X)}}[\alpha(\gamma)^2]$. For a summary of such aggregations see (Garg et al., 2024). We illustrate Definition 5 in Figure 4.

5.1 Distribution Calibration Implies Γ-Calibration

Distribution calibration with respect to a property Γ almost immediately implies Γ-calibration. The argument naturally seems correct, but it requires careful treatment of convex combination of distributions on which distributional predictions with equal property values have been made. For the sake of readability, we restrict Proposition 6 to finite $\mathcal{Y}$. The extension to more general $\mathcal{Y}$ requires more technical machinery. We give the statement in the appendix (Proposition 20).

Proposition 6 (Distribution Calibration with Respect to Γ implies Γ-Calibration)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a predictor which is distribution calibrated with respect to Γ. Then, $\Gamma \circ f$ is Γ-calibrated.*

Proof We have to show that for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$\Gamma(D_{Y|\Gamma \circ f(X)=\gamma}) = \gamma. \tag{1}$$

Let $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$ be arbitrary. By distribution calibration with respect to Γ,

$$D_{Y|\Gamma \circ f(X)=\gamma} = \mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma].$$

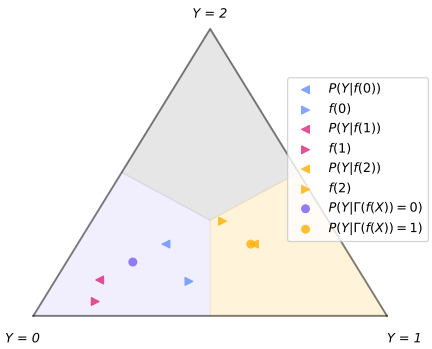
Lemma 38 applied to D, f and $A \coloneqq \{x \in \mathcal{X}\colon \Gamma \circ f(x) = \gamma\}$ gives,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) = \gamma,$$

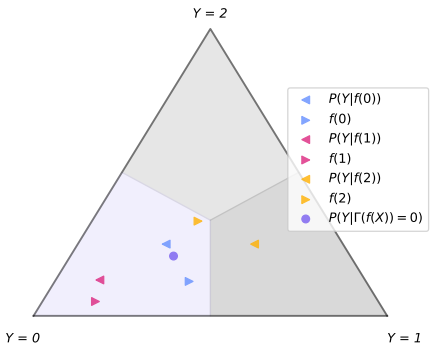
hence Equation 1 follows. ■

The implication can, in general, *not* be extended to a bi-implication. Figure 2b and Figure 4a are illustrations for the same predictor, which is approximately distribution calibrated but perfectly Γ-calibrated. The intuition is that Γ-calibration makes the grid for testing calibration coarser.

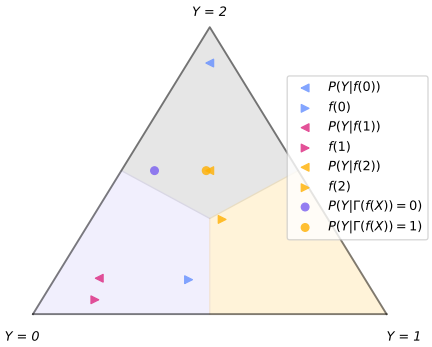
The above Proposition 6 holds for arbitrary properties with convex level sets. In the approximate case, we show that the implication continues to hold if the property Γ is Lipschitz continuous (Appendix B.1). Note that discrete properties Γ don't fulfill this condition, because a slight mismatch of the average predictions can lead to a catastrophic mismatch in the corresponding Γ-values.



(a) Perfect Γ-calibration.



(b) Perfect Γ-calibration II.



(c) Approximate Γ-calibration.

Figure 4: Illustration of Γ-calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $\mathcal{X} = \{0, 1, 2\}$. We define $\Gamma(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ are colored respectively drawn in different shades of gray. The left-directing markers denote the true conditional distribution for different choices of $x \in \mathcal{X}$. We assume that f is a distributional predictor (right-directing markers) which is then fed into Γ. The dots denote the true outcome distribution conditioned on a value of $\Gamma \circ f$. If the dot is in the level set of the same color, then the prediction is perfectly Γ-calibrated. Note that Figure 4a uses the same predictions and outcome distributions as Figure 2b showing that predictions could be perfect Γ-calibrated while only being approximately distribution calibrated with respect to Γ. Figure 4b shows that Γ-calibration does not necessarily require the true conditional distributions to all live in the correct level set. Only the predictor f is changed from Figure 4a to Figure 4b. Figure 4c then changes the true conditional distribution compared to Figure 4a but fixes f, which lead to the violation of the perfect Γ-calibration constraint.

5.2 Γ-Calibration is Equivalent to Low Swap Regret

Property calibration can be equivalently reformulated for loss (respectively identification) functions, if the property is elicitable (respectively identifiable).[13]

Proposition 7 (Perfect Γ-Calibration via Loss Function or Identification Function)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be Γ-calibrated on D.*

(a) Suppose Γ is elicitable with ℓ, then f is Γ-calibrated on D if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$\mathbb{E}_D\left[\ell(Y,\gamma)|f(X)=\gamma\right] - \min_{\hat{\gamma}\in \operatorname{im}\Gamma} \mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma\right] = 0.$$

(b) Suppose Γ is identifiable with V, then f is Γ-calibrated on D if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$\mathbb{E}_D\left[V(Y,\gamma)|f(X)=\gamma\right] = 0.$$

Proof By definition of elicitability and identifiability. ■

Statement (a) is essentially a "no swap regret" statement. The predictor f is compared with the best prediction $\hat{\gamma}^*$ on the sets on which the predictor f predicted a value γ. If statement (a) is not perfectly but approximately fulfilled, we call it "low swap regret". No swap regret is different from precise loss estimation as we argue in Section 6. The no swap regret transfers to the approximate case, i.e., low swap regret. For that, certain regularity assumptions on the property Γ and its identification function V are needed. In Appendix B.2 we argue for the general correspondence and link it to the work on swap multicalibration and swap-agnostic learning (Gopalan et al., 2024b).

5.3 Γ-Calibration is Inherited

There is a commonplace gap between *optimization-level* Γ and *decision-level* properties Φ which trades off continuity for a granularity unnecessary for decision-making. Given this gap, we have to ask if self-realization with respect to Γ extends to a related Φ. In the sequel, we show that self-realization, as Γ-calibration, is inherited by *refined* properties (Definition 3).[14] That way, calibrated forecasts for optimization-level properties provide calibrated estimates for decision-level properties. In the language of decision making, Γ-calibration guarantees low swap regret for downstream decision makers.[15]

Proposition 8 (Γ-Calibration is Inherited by Refined Properties)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property, D a regular data distribution on $\mathcal{X} \times \mathcal{Y}$ with respect to $\mathcal{P}$. Suppose that $\mathcal{Y}$ is finite*

13. Already in (Davis, 2016) the terms "calibration" and "identification" got linked to each other. However, their Definition 4.2 of calibration is largely disconnected from the use of the term in current machine learning literature.
14. The argument seems more involved then necessary. This is a consequence of carefully arguing about convex combinations of outcome distributions analogous to Proposition 6.
15. This provides a major motivation for self-realization in the first place (Noarov and Roth, 2024).

and f is a Γ-predictor which is Γ-calibrated on D. Suppose, furthermore, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. Then, for every property Φ with convex level sets and which is refined by Γ, the Φ-predictor $\phi \circ f$, where ϕ is defined following Definition 3, is Φ-calibrated on D.

Proof We have to show that,

$$\Phi(D_{Y|\phi\circ f(X)=v}) = \phi \circ \Gamma(D_{Y|\phi\circ f(X)=v}) = v,$$

for all $v \in \operatorname{supp} D_{\phi\circ f(X)}$. Fix $v \in \operatorname{supp} D_{\phi\circ f(X)}$ for the following. Let us define the probabilistic predictor $h\colon \mathcal{X} \to \mathcal{P}$ where $x \mapsto D_{Y|f(X)=f(x)}$. The mapping is measurable by definition, as f is measurable by assumption. Let $A \coloneqq \{x \in X\colon \phi \circ f(x) = v\}$. Since, f is Γ-calibrated, and for all $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$, it holds,

$$\Gamma(h(x)) = \Gamma(D_{Y|f(X)=f(x)}) = f(x)$$

for all $x \in A$. Hence, in particular,

$$\Phi(h(x)) = v,$$

for all $x \in A$. By the law of total expectation, we obtain, for all $y \in \mathcal{Y}$,

$$\begin{aligned} D_{Y|\phi\circ f(X)=v}(Y = y) &= \mathbb{E}_D[[\![Y = y]\!]|\phi(f(X)) = v] \\ &= \mathbb{E}_{G\sim D_{f(X)}}\left[\mathbb{E}_D[[\![Y = y]\!]|f(X) = G]|\phi(G) = v\right] \\ &= \mathbb{E}_{G\sim D_{f(X)}}\left[\mathcal{D}_{Y|f(X)=G}(Y = y)|\phi(G) = v\right]. \end{aligned}$$

It follows by definition of h, A and because $\mathcal{Y}$ is finite, $D_{Y|\phi\circ f(X)=v} = \mathbb{E}_{D_X}[h(X)|X \in A]$. Applying Lemma 38 to h, D and A gives,

$$\Phi(D_{Y|\phi\circ f(X)=v}) = \Phi(\mathbb{E}_{\sim D_X}[h(X)|X \in A]) = v.$$

■

The above statement can be extended to more general $\mathcal{Y}$ beyond finiteness (Proposition 21).

Given that the Γ-predictor f is *not* perfectly calibrated, we show two approaches to elaborate the inheritance statement for approximately calibrated f in the appendix. First, we show that if Γ is Lipschitz, a relaxed inheritance statement continues to hold (Proposition 28). b) Second, we argue that if the decision-level property Φ is elicited by a Lipschitz loss function ℓ, a relatively weak assumption, then approximate Γ-calibration implies low swap regret with respect to this loss function ℓ (Proposition 30). In particular, the second approach is contextualized within the paradigm of "omniprediction" (Gopalan et al., 2022b) (Section B.3).

The inheritance of self-realization is central to the rationale of using calibration. It provides low swap regret guarantees for a large class of loss functions (Proposition 30). Each such loss function corresponds to an elicitable property, a "decision-level" property. For instance, this decision-level property can model a person who is about to decide whether to take or not to take their umbrella with them given a probabilistic forecast on rain. However, self-realization can have largely decoupled effects on the usefulness of the predictions for the "decision-level" properties, respectively the decision makers. A calibrated prediction can be more useful to one than to another decision maker (Appendix G).

6. Calibration as Precise Loss Estimation

In contrast with the account laid out above, Zhao et al. (2021) motivate calibration differently. They consider a predictor to be calibrated, if it can precisely[16] estimate the average loss incurred to the individuals. Formally, the intuition can be captured by a type of loss outcome indistinguishability (Gopalan et al., 2022a) following (Zhao et al., 2021).

Definition 9 (Decision Calibration (Modified from Zhao et al. (2021))). *Let Γ be an elicitable property and $\mathcal{L}$ be a set of $\mathcal{P}$-consistent loss functions $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ with respect to Γ. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$ and $f\colon \mathcal{X} \to \mathcal{P}$ a distribution predictor. The predictor f is* β-approximate decision calibrated *for $\mathcal{L}$, if for all $\ell \in \mathcal{L}$,*

$$\left|\mathbb{E}_{(X,Y)\sim D}[\ell(Y, \Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y}, \Gamma(f(X))]]\right| \leq \beta.$$

In the original definition (Zhao et al., 2021, Definition 2) the loss function ℓ and the elicited property Γ can be independent of each other. For instance, the loss function might penalize deviations of $f(X)$ from the mean, while Γ maps $f(X)$ to the median. That is a semantic mismatch. We were not able to identify a convincing justification for this choice, hence we neglect this degree of freedom for our purposes. Hence, we suggest a definition which generalizes "decision outcome indistinguishability" beyond the binary (Gopalan et al., 2022a).

In our definition we follow the presentation by Fröhlich and Williamson (2024). They back the intuition of "calibration via precise loss estimation" from an actuarial point of view referring to the ideal of *actuarial fairness* (Arrow, 1978):

> The forecaster should offer actuarially fair insurance for the uncertain loss [...]. Actuarial fairness here means that in the long run, the forecaster neither loses nor profits from offering insurance under the data model. (Fröhlich and Williamson, 2024, p. 12)

In other words, both parties get a fair deal. The forecaster could, without getting bankrupt, ask for a certain premium to all its decision makers based on the forecast and in turn cover all losses the decision maker might encounter because of the uncertain outcomes. Decision calibration fits to the narrative of "trustworthiness" via precise loss estimates encouraged by Kirchhof et al. (2024) and Yoo and Kweon (2019) and which can be traced back to at least (Rukhin, 1988). Note in those papers precise loss estimation is usually desired for every individual. In contrast, we focus on average precise loss estimation here. Decision calibration enforced on sub-groups based on input X bridge between the two extremes: individual precise loss estimation versus average precise loss estimation (cf. Section 7).

Decision calibration, like all other so far named notions of calibration, grows on the ground of binary vanilla calibration (cf. Proposition 11). We show that on binary outcome sets decision calibration with respect to the set of *simple loss functions*, also known as "cost-weighted misclassification losses", is equivalent to binary vanilla calibration. Simple loss functions play a key role in defining the spectrum of proper loss functions for binary outcome sets, (Schervish, 1989) (for a modern proof see (Reid and Williamson, 2011, Theorem 16) and an independent rediscovery (Kleinberg et al., 2023)).

16. We deliberately refer to "precise" loss estimates as we focus on the internal consistency of the loss estimates and not the actual closeness to the truth (accuracy) (Everitt, 2002, p. 295).

Definition 10 (Simple Loss Function)**.** *Let $q \in [0,1]$. A loss function $\ell_q \colon \{0,1\} \times \{a,b\} \to \mathbb{R}$ is called* simple *if it has the form,*

$$\ell_q(y,c) = q[\![y=0, c=a]\!] + (1-q)[\![y=1, c=b]\!].$$

The loss ℓ_q elicits the simple binary property,

$$\Gamma_q \colon \Delta(\mathcal{Y}) \to \{a,b\}; P \mapsto \begin{cases} a \text{ if } P(Y=1) > q \\ b \text{ if } P(Y=1) \leq q. \end{cases}$$

We overload the notation of a simple loss function with its proper loss cousin: $\ell_q \colon \{0,1\} \times [0,1] \to \mathbb{R}$, $\ell_q(y,p) = \ell_q(y, \Gamma_q(p))$.

The equivalence of decision calibration and binary vanilla calibration has already been discussed in (Zhao et al., 2021, Theorem 1). However, their proof is hard to follow and makes heavy use of the independent choice of the decision function and the loss function, which we hesitated to commit to (see discussion below Definition 9). Hence, our statement requires a weaker definition of decision calibration. On the other hand, our result only holds on binary outcome sets and requires predictors with finitely many output values, arguably a mild assumption for practical settings.[17]

Proposition 11 (Decision Calibration Equivalent to Vanilla Calibration for Binary Outcome Set)**.** *Let $\mathcal{Y} = \{0,1\}$ and suppose $\mathcal{X}$ is arbitrary. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f \colon \mathcal{X} \to [0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y=1$, and suppose that $|\mathrm{im} f| < \infty$ and $D_X(f(x)=v) > 0$ for all $v \in \mathrm{im} f$. Then, f is calibrated if, and only if, f is decision calibrated with respect to all simple loss functions $\{\ell_q \colon q \in [0,1]\}$.*

Proof First, we consider two different cases. Let $f_{\inf} \coloneqq \inf_{\gamma \in \mathrm{im} f} \gamma$.

(a) If $f_{\inf} > 0$, then Lemma 41 applies, hence f *matches the averages*, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$.

(b) If $f_{\inf} = 0$, then Lemma 42 applies because $\mathrm{im} f$ is finite, hence f *matches the averages*, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$.

Lemma 40 can be used to show that for all $q \in [0,1]$, $\mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) \leq q] = 0$ and $\mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) > q] = 0$. It remains to show that for all $v \in \mathrm{im} f$, $\mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) = v] = 0$. If $v = 0$, this is already given. For every $v \in \mathrm{im} f \setminus \{0\}$ there exists $v' \in \mathrm{im} f \cup \{0\}$ such that $v' < v$ (because $\mathrm{im} f$ is finite). We have,

$$\begin{aligned} &\mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) \leq v] \\ &= P_D(f(X) \leq v') \mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) \leq v'] \\ &\quad + P_D(f(X) = v) \mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) = v] \\ &= P_D(f(X) = v) \mathbb{E}_{(X,Y)\sim D}[Y - f(X) | f(X) = v] = 0, \end{aligned}$$

17. For an intuitive argument why Proposition 11 should hold see the example provided in (Höltgen, 2024, Section 2.2).

which gives the desired result. ■

In light of Proposition 11, the "precise loss estimation"-account is another generalization of binary vanilla calibration whose semantic focus is distinct from the self-realization discussed in § 5. Furthermore, precise loss estimation via decision calibration is also a child of the general notion of distribution calibration.

6.1 Distribution Calibration Implies Decision Calibration

Rather unsurprisingly, distribution calibration (Definition 2) implies decision calibration (Definition 9). This holds not only in the perfect case but as well in the case of approximate calibration if the loss function corresponding to the property of interest is bounded (Appendix B.4). Proposition 12 (respectively the approximate version Proposition 31) has been proven in (Zhao et al., 2021, Proposition 2). However, we were not able to follow all steps of the proofs. For this reason we provide a full argument here.

Proposition 12 (Distribution Calibration Implies Decision Calibration)**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor which is distribution calibrated with respect to Γ on D. Then, f is decision calibrated with respect to $\mathcal{L}_\Gamma \coloneqq \{\ell\colon \ell$ is $\mathcal{P}$-consistent loss function for $\Gamma\}$.*

Proof The predictor f is *distribution calibrated with respect to* Γ on D if for every $\gamma \in \operatorname{supp} D_{\Gamma(f(X))}$,

$$D_{Y|\Gamma\circ f(X)=\gamma}(Y = y) = \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma \circ f(X) = \gamma], \quad \forall y \in \mathcal{Y}, \tag{2}$$

where $f_y(x) \in [0, 1]$ denotes the y-component of the prediction $f(x) \in \Delta(\mathcal{Y})$ for $x \in \mathcal{X}$.

Let $\ell \in \mathcal{L}_\Gamma$ be arbitrary. For every $\gamma \in \operatorname{supp} D_{\Gamma(f(X))}$, we have,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma \circ f(X) = \gamma\right] \\
&= \mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma)|\Gamma \circ f(X) = \gamma\right] - \mathbb{E}_{(X,Y)\sim D}\left[\mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma \circ f(X) = \gamma\right] \\
&= \sum_{y\in\mathcal{Y}} D_{Y|\Gamma\circ f(X)=\gamma}(Y = y)\ell(y,\gamma) - \mathbb{E}_{X\sim D_X}\left[\sum_{y\in\mathcal{Y}} f_y(X)\ell(y,\gamma) \,\middle|\, \Gamma \circ f(X) = \gamma\right] \\
&= \sum_{y\in\mathcal{Y}} \Big(D_{Y|\Gamma\circ f(X)=\gamma}(Y = y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma \circ f(X) = \gamma]\Big)\,\ell(y,\gamma) \\
&= 0,
\end{aligned}$$

by Equation (2). Hence, it follows,

$$\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\Gamma(f(X)))]\right] = 0.$$

■

Is the reverse direction true? Proposition 11 provides an affirmative answer in the case of binary outcomes. However, the proof is not immediately extendable beyond the binary. The question remains open for future work.

6.2 Precise Bayes Risk Estimation

Analogous to self-realization we obtain a clearer picture of the purpose of decision calibration when rewriting in terms of properties. To this end, we introduce Bayes pairs. A Bayes pair unifies the minimizer (elicited property) and the minimization value (Bayes risk) of a loss function.[18] Hence, following the account of precise loss estimation the Bayes risk is of central importance here. Being calibrated in the sense of precise loss estimation, is then tantamount to precise estimation of the Bayes risk.

Definition 13 (Bayes Pair (Embrechts et al., 2021))**.** *Let $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ be a loss function. The property pair (Φ_ℓ, Θ_ℓ) is called a Bayes pair, if, for all $P \in \mathcal{P}$,*[19]

$$\Phi_\ell(P) = \arg\min_{\phi \in \mathcal{R}} \mathbb{E}_{Y\sim P}[\ell(Y, \phi)],$$
$$\Theta_\ell(P) = \min_{\phi \in \mathcal{R}} \mathbb{E}_{Y\sim P}[\ell(Y, \phi)].$$

By definition, Φ_ℓ is elicitable. That the Bayes risk Θ_ℓ is elicitable on level sets of Φ_ℓ is less obvious. To see this, observe that there exists a $\Phi_\ell^{-1}(\phi)$-consistent loss function for Θ_ℓ for every $\phi \in \mathrm{im}\Phi_\ell$ (Embrechts et al., 2021). We proceed by defining precise Bayes risk estimators.

Definition 14 (Precise Bayes Risk Estimator)**.** *Let $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ be a loss function with corresponding Bayes pair (Φ_ℓ, Θ_ℓ). Let D be a regular data distribution. A (Φ_ℓ, Θ_ℓ)-predictor $f = (g, h)$ is a* β-precise Bayes risk estimator *if,*

$$|\mathbb{E}_D[\ell(Y, g(X)) - h(X)]| \leq \beta.$$

This definition is extremely weak. The predictor h estimates the loss of g but only has to fit it on average. Additionally, there is *no* demand placed on the quality of g. The definition can be significantly strengthened if precise Bayes risk estimation is demanded on subgroups or individuals (cf. Section 7). Note that such Bayes risk predictors exist in current literature. For instance, they are called "uncertainty module" (Kirchhof et al., 2024) or "loss prediction module" (Yoo and Kweon, 2019). It follows rather immediately that a decision calibrated full probability predictor is a precise Bayes risk estimator.

Proposition 15 (Decision Calibration Implies Precise Bayes Risk Estimation)**.** *Let $\mathcal{L}$ be a set of $\mathcal{P}$-consistent loss functions $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ with corresponding Bayes pairs (Φ_ℓ, Θ_ℓ). Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$ and $f\colon \mathcal{X} \to \mathcal{P}$ a β-approximate decision calibrated predictor for $\mathcal{L}$. Then, for all $\ell \in \mathcal{L}$ the predictor $(\Phi_\ell \circ f, \Theta_\ell \circ f)$ is a β-precise Bayes risk estimator.*

Proof We can rewrite the criterion of decision calibration as,

$$\left|\mathbb{E}_D[\ell(Y, \Phi_\ell(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y}, \Phi_\ell(f(X)))]]\right| = |\mathbb{E}_D[\ell(Y, \Phi_\ell(f(X))) - \Theta_\ell(f(X))]| \leq \beta. \tag{3}$$

■

18. The importance of eliciting Bayes pairs is also studied by (Fissler and Ziegel, 2016; Frongillo and Kash, 2021; Finocchiaro et al., 2021).
19. Under the restriction that $\mathcal{R}$ is compact both properties are guaranteed to be measurable (Aliprantis and Border, 2006, Theorem 18.19).

6.3 When Self-Realization Implies Precise Loss Estimation

In the sections before we explored the landscape of calibration as self-realization and as precise loss estimation. We introduced two prototypical notions of calibration using properties for each of the two worlds: property calibration and precise Bayes risk estimation. In the following section, we present several results and examples which shed light on the complicated relationship between the notions.

If we have access to a predictor which predicts a Bayes pair and is property-calibrated with respect to the involved Bayes risk property, then the predictor is a precise Bayes risk estimator (Proposition 16). That is, property calibration for the property (Γ, Θ) implies precise Bayes risk estimation.

Proposition 16 (Self-Realization Implies Precise Loss Estimation)**.** *Let (Φ_ℓ, Θ_ℓ) be a Bayes pair corresponding to a loss function ℓ and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. If a (Φ_ℓ, Θ_ℓ)-predictor $f = (g, h)$ is $\alpha(\phi, \theta)$-approximate (Φ_ℓ, Θ_ℓ)-calibrated, then f is a α-precise Bayes risk estimator, where $\alpha \coloneqq \mathbb{E}_{(\phi,\theta)\sim Q}[|\alpha(\phi, \theta)|]$ and $Q \coloneqq D_{(g(X),h(X))}$*

Proof

$$
\begin{aligned}
&|\mathbb{E}_D[\ell(Y, g(X)) - h(X)]| \\
&= |\mathbb{E}_{(\phi,\theta)\sim Q}[\mathbb{E}_{(X,Y)\sim D|g(X)=\phi,h(X)=\theta}[\ell(Y, \phi)] - \theta]| \\
&\leq \mathbb{E}_{(\phi,\theta)\sim Q}[|\Theta_\ell(D_{Y|g(X)=\phi,h(X)=\theta}) - \theta|] \\
&\leq \mathbb{E}_{(\phi,\theta)\sim Q}[|\alpha(\phi, \theta)|] = \alpha.
\end{aligned}
$$

■

The implication cannot be generally extended into a bi-implication as the following example shows. In this example, we construct a Bayes pair predictor, i.e., a predictor which predicts the first and second component of a Bayes pair, which estimates the Bayes risk precisely. However, this predictor is not calibrated with respect to the first or second component of the Bayes pair. A instantiation of this example would be a $(\mathbb{M}, \mathbb{V})$-predictor, mean and variance predictor, which precisely estimates the Bayes risk, but is not $(\mathbb{M}, \mathbb{V})$-calibrated, nor is the first component of the predictor $\mathbb{M}$-calibrated, i.e., mean-calibrated, or the second component $\mathbb{V}$-calibrated, i.e., variance-calibrated. In that sense, self-realization is "stronger" than precise loss estimation.

Example 1 (On the Necessity of Self-Realization for Precise Loss Estimation)**.** *The following example is based on an arbitrarily chosen Bayes pair. The idea is that a predictor of the Bayes pair which swaps the elicited value, i.e., the first component of the Bayes pair, can still precisely estimate the Bayes risk. In other words, the predictor is biased, but still precisely estimate its error.*

Let (Φ_ℓ, Θ_ℓ) be the Bayes pair corresponding to the loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$. Let (Φ_ℓ, Θ_ℓ) be the Bayes pair corresponding to the loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Suppose that there exist $A \subseteq \mathcal{X}$, such that $1 > D_X(A) > 0$ and

$$
\begin{aligned}
\Phi_\ell(D_{Y|X\in A}) &= \phi_A, & \Theta_\ell(D_{Y|X\in A}) &= \mathbb{E}_{D_Y}[\ell(Y, \phi_A)|X \in A], \\
\Phi_\ell(D_{Y|X\in \overline{A}}) &= \phi_{\overline{A}}, & \Theta_\ell(D_{Y|X\in \overline{A}}) &= \mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}})|X \in \overline{A}],
\end{aligned}
$$

where $\phi_A \neq \phi_{\overline{A}}$.

Let (f, g) *be a stacked predictor of the Bayes pair* (Φ_ℓ, Θ_ℓ). *Suppose that,*

$$f(x) = \begin{cases} \phi_{\overline{A}} \text{ for } x \in A \\ \phi_A \text{ for } x \in \overline{A}. \end{cases},$$

$$g(x) = \begin{cases} \mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}})|X \in A] & \text{for } x \in A \\ \mathbb{E}_{D_Y}[\ell(Y, \phi_A)|X \in \overline{A}] & \text{for } x \in \overline{A}. \end{cases}.$$

We can directly show that (f, g) *is a precise loss estimator,*

$$\begin{aligned} &\mathbb{E}_D[\ell(Y, f(X)) - g(X)] \\ &= D_X(A)\mathbb{E}_{D_Y}[\ell(Y, f(X)) - g(X)|X \in A] + D_X(\overline{A})\mathbb{E}_{D_Y}[\ell(Y, f(X)) - g(X)|X \in \overline{A}] \\ &= D_X(A)\mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}}) - g(X)|X \in A] + D_X(\overline{A})\mathbb{E}_{D_Y}[\ell(Y, \phi_A) - g(X)|X \in \overline{A}] \\ &= 0. \end{aligned}$$

On the other hand, f *is clearly not* Φ_ℓ*-calibrated, because,*

$$\begin{aligned} \Phi_\ell(D_{Y|f(X)=\phi_{\overline{A}}}) &= \Phi_\ell(D_{Y|X\in A}) = \phi_A \neq \phi_{\overline{A}}, \\ \Phi_\ell(D_{Y|f(X)=\phi_A}) &= \Phi_\ell(D_{Y|X\in\overline{A}}) = \phi_{\overline{A}} \neq \phi_A. \end{aligned}$$

Furthermore, g *is not* Θ_ℓ*-calibrated, because*

$$\begin{aligned} \Theta_\ell(D_{Y|g(x)=\mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X\in A]}) &= \Theta_\ell(D_{Y|X\in A}) \\ &= \mathbb{E}_{D_Y}[\ell(Y, \phi_A)|X \in A] < \mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}})|X \in A], \end{aligned}$$

and,

$$\begin{aligned} \Theta_\ell(D_{Y|g(x)=\mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X\in\overline{A}]}) &= \Theta_\ell(D_{Y|X\in\overline{A}}) \\ &= \mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}})|X \in \overline{A}] < \mathbb{E}_{D_Y}[\ell(Y, \phi_A)|X \in \overline{A}], \end{aligned}$$

by (strict) consistency of the loss function ℓ.

However, precise loss estimation always requires one to predict not only the property of interest but as well the corresponding Bayes risk. For instance, a pure mean predictor f as in Example 1 can be perfectly calibrated but still misses the information for being a precise Bayes risk estimator. In that sense, Γ-calibration and precise Bayes risk estimation are incommensurable as they are properties of different predictors. There is a type mismatch. This explains why distribution calibration, which requires full prediction estimates, subsumes both: all properties of a distribution including the Bayes risks are refined by the full distribution.

Nevertheless, Example 1 can be strengthened. If $\mathcal{Y}$ is binary, then the mean predictor identifies the full distribution. But even then, there exist predictors which are perfectly decision calibrated with respect to the squared loss (which implies precise Bayes risk estimation cf. Proposition 15), but still are not mean calibrated. We argue in the following, that decision calibration is *too* weak to imply property calibration (cf. Proposition 11). More

concretely, we show in the following lemma that for any continuous loss function which elicits the mean there is a constant mean-predictor which is decision calibrated with respect to the loss function, independent of the data distribution. Hence, depending on the data distribution the constant mean-predictor is not mean-calibrated, as we argue afterwards.

Lemma 17 (Existence of Constant Decision Calibrated Mean Predictor)**.** *Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \mathbb{R}$. For any continuous proper loss function $\ell\colon \{0,1\} \times [0,1] \to \mathbb{R}$ which elicits the mean on $\mathcal{Y}$, there is a (constant) mean predictor $f\colon \mathcal{X} \to [0,1]$ which is decision calibrated*[20] *for $\{\ell\}$ on every data distribution D on $\mathcal{X} \times \mathcal{Y}$.*

Proof First, we argue that there exist $q^* \in [0,1]$ such that $\ell(0, q^*) = \ell(1, q^*)$. This is true, because of properness, for all $q \in [0,1]$,

$$\ell(0,0) \leq \ell(0,q)$$
$$\ell(1,1) \leq \ell(1,q).$$

Which implies that $h\colon q \mapsto \ell(1,q) - \ell(0,q)$, fulfills,

$$h(0) = \ell(1,0) - \ell(0,0) \geq 0$$
$$h(1) = \ell(1,1) - \ell(0,1) \leq 0.$$

By continuity of h, because of continuity of ℓ, it follows by the intermediate value theorem, that there exists $q^* \in [0,1]$ such that $h(q^*) = 0$.
Second, let us rewrite the decision calibration condition, for data distribution D,

$$\begin{aligned}&\mathbb{E}_{(X,Y)\sim D}[\ell(Y, f(X) - \mathbb{E}_{\hat{Y}\sim(1-f(X),f(X))}[\ell(\hat{Y}, f(X)]]\\ &= \mathbb{E}_{X\sim D_X}[\mathbb{E}_{Y\sim D_Y}[\ell(Y, f(x)|X = x] - \mathbb{E}_{\hat{Y}\sim(1-f(x),f(x))}[\ell(\hat{Y}, f(x)|X = x]].\end{aligned}$$

Define $p(x) \coloneqq D_{Y|X=x}(\{1\})$ and focus on the inner part of the expectation conditioned on $X = x$,

$$\mathbb{E}_{Y\sim D_Y}[\ell(Y, f(x))|X = x] = p(x)\left(\ell(1, f(x)) - \ell(0, f(x))\right) + \ell(0, f(x))$$
$$\mathbb{E}_{Y\sim(1-f(X),f(X))}[\ell(Y, f(x))|X = x] = f(x)\left(\ell(1, f(x)) - \ell(0, f(x))\right) + \ell(0, f(x)).$$

That is, the inner difference is,

$$\left(p(x) - f(x)\right)\left(\ell(1, f(x)) - \ell(0, f(x))\right),$$

which is 0 if $f(x) = q^*$. This proves the statement. ■

It is now a matter of a simple construction for a specific data distribution D to show that a mean-predictor is decision calibrated but not mean-calibrated.

Example 2 (Decision-calibrated Mean-Predictor Which is Not Mean-Calibrated)**.** *Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X} = \mathbb{R}$. Pick $f\colon \mathcal{X} \to [0,1]$ to be a constant predictor $f(x) = q^*$ which is decision calibrated for the continuous proper loss function $\ell\colon \{0,1\} \times [0,1] \to \mathbb{R}$. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}_{Y\sim D_Y}[Y] = q \neq q^*$. Then, clearly f is* not *mean-calibrated,*

$$\mathbb{M}(D_{Y|f(X)=q^*}) = q \neq q^*.$$

20. We abuse the term decision calibration here a bit. Decision calibration following Definition 9 is defined for probabilistic predictors. In our present case, the predictor is a mean predictor. However, the mean fully identifies a distribution because we are in a binary setting.

7. What about groups?

Certain frameworks like fair machine learning demand calibration on subgroups defined through the input space $\mathcal{X}$. Hence, one can understand calibration in fair machine learning, as asking for "equally good self-realization of predictions on sensitive subgroups" or "equally precise loss estimation on sensitive subgroups". The notions of calibration we considered can easily be extended to subgroups. For instance, by taking a supremum over subgroups in Definition 5 one recovers the notion of multi-calibration used in (Noarov and Roth, 2024). Respectively, a precise Bayes risk estimator following Definition 14 can be strengthened by checking for precise Bayes risk estimation on all relevant subgroups. Note that the choice of groups on which a certain notion of calibration ought to hold is *independent* of the choice of the notion itself.

In the extreme case, those subgroups could be enforced to the level of individuality (Zhao et al., 2020; Luo et al., 2022). Then, "trustworthiness" meets "usefulness" again. The Bayes optimal predictor is the only predictor which achieves "trustworthiness" on all individuals. In particular, it is possible to equalize "trustworthiness" on arbitrary subgroups without losing on "usefulness". However, it is not possible to equalize "usefulness" on arbitrary subgroups without losing on "trustworthiness" (Barocas et al., 2023, Proposition 4 & 5).

Concluding, all considerations made on the semantics of the three different types of calibration considered in this work plus their formal relationships directly transfer to the fairness setting with multiple groups.

8. Conclusion

This work started with the promise to provide semantical structure in the chaotic world of calibration notions. We consider as our main contribution to semantically and formally separate two concerns calibration wants to solve: self-realization and precise loss estimation. Both accounts differently motivate a certain type of calibration. However, they can be bridged by distribution calibration.

How do existing notions relate to the three different types? Table 1 and Table 2 summarize a list of existing notions and their categorization in the three different types. Note that we make a distinction between formal strict derivatives, i.e., our suggested notion generalizes the notion in the list, and definitions which follow the type of definition more broadly.

What is the learning for a practitioner? We have *not* provided any new algorithm nor have we suggested *the* notion of calibration. But, our work shapes the debate on what is the right notion. It highlights implications (see Figure 1 and Figure 6), incommensurabilities (see Section 6.3) and equivalences (Proposition 33, Proposition 23 and Proposition 11). If those relationships are ignored, this may lead to unforeseen, unintentional consequences or may unnecessarily complicate a prediction problem[21]. Our work helps to define a set of questions which are relevant to formalizing a real-world prediction problem. What are the predictions we can or should get? What and who are the agents which use the predictions? Is one of the two accounts laid out desired as quality criterion of the predictions? Finally, our work advocates leaving the formal distinctions between evaluation metrics behind (Proposition 26)

21. For instance, in the example given in Section 2 any calibration beyond the ranking is unnecessary, since the policy is bound to the rankings only (Perdomo et al., 2023).

Formally Strict Derivative	Name	Reference
Distribution Calibration	class-wise calibration	(Kull et al., 2019)
	confidence calibration	(Guo et al., 2017)
	event-conditional unbiasedness (without groupings via $\mathcal{X}$)	(Noarov et al., 2023)
Γ-Calibration	Γ-calibration (without groupings via $\mathcal{X}$)	(Noarov and Roth, 2023)
	T-calibration	(Gneiting and Resin, 2023)
	quantile calibration	(Kuleshov and Deshpande, 2022)
Decision Calibration	decision calibration	(Zhao et al., 2021)
	decision outcome indistinguishability	(Gopalan et al., 2022a)

Table 1: Categorization of existing notions of calibration in the three different types following by which type the notion is formally generalized. This table does not claim to contain a complete list of all definitions of calibration.

Comparable Type of Definition	Name	Reference
Distribution Calibration	distribution calibration	(Song et al., 2019)
	calibration safety	(Grünwald, 2018)
Γ-Calibration	marginal coverage calibration	(Räth and Ludwig, 2025)
	calibration safety	(Grünwald, 2018)
Decision Calibration	loss outcome indistinguishability	(Gopalan et al., 2022a)
	IP-calibration (without groupings via $\mathcal{X}$)	(Fröhlich and Williamson, 2024)

Table 2: Categorization of existing notions of calibration in the three different types following a broader comparability of the account. This table does not claim to contain a complete list of all definitions of calibration.

and re-focus on the actual purpose of evaluation: estimation of usefulness, self-realization or guarantees on the estimation of usefulness. Particularly, calibration is only a part of this bigger picture.

9. Acknowledgements

The authors are very grateful to Benedikt Höltgen and Rajeev Verma for feedback on an earlier draft of this work. Thanks to the International Max Planck Research School for Intelligent System (IMPRS-IS) for supporting Rabanus Derr. Thanks to James Bailie for giving a reason to visit Cambridge, MA where Jessie and Rabanus met the first time. Part of the research happened when Rabanus was visiting Aaron Roth supported by a DAAD IFI-Stipend. Rabanus Derr and Robert C. Williamson were funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645; they were also supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center.

Jessie Finocchiaro was supported in part by the National Science Foundation under Award No. 2202898. We thank the reviewers for helpful and constructive feedback.

Appendix A. Distribution Calibration for General $\mathcal{Y}$

Our definition of distribution calibration (Definition 2) with respect to a property Γ asks for $\mathcal{Y}$ to be finite. This is *not* a logically required part of the definition and rather simplifies the presentation. We can define distribution calibration on arbitrary $\mathcal{Y}$ by referring to measurable sets $A \in \mathcal{B}(\mathcal{Y})$. We restrict all definitions and statements in this section to the perfect case.

Definition 18 (Distribution Calibration with Respect to Γ on General $\mathcal{Y}$)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, regular wrt. $\mathcal{P}$. Suppose that $\mathcal{P} \subseteq W_{TV}$ is a subset of a Banach space equipped with the total variation distance. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. The predictor f is* distribution calibrated with respect to Γ *on D if for every $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$ and $A \in \mathcal{B}(\mathcal{Y})$,*

$$D_{Y|\Gamma \circ f(X)=\gamma}(A) = \mathbb{E}_D[f_A(X)|\Gamma \circ f(X) = \gamma]. \tag{4}$$

In the following we extend Proposition 4, Proposition 6, Proposition 8 and Proposition 12 to general $\mathcal{Y}$ using the new definition of distribution calibration. We summarize the extension in Table 3. All statements are given for perfect calibration. Note that with the additional statements of this section the implications in Figure 1 hold for general $\mathcal{Y}$ (Figure 5). To prove the implication structure we need to extend our technical machinery a bit. The technical ideas are heavily inspired by (Noarov and Roth, 2023).

Statement	Finite $\mathcal{Y}$	General $\mathcal{Y}$
Distribution calibration is inherited	Proposition 4	Proposition 19
Distribution calibration implies property calibration	Proposition 6	Proposition 20
Property calibration is inherited	Proposition 8	Proposition 21
Distribution calibration implies decision calibration	Proposition 12	Proposition 22

Table 3: Summary of extended statements from $\mathcal{Y}$ to general $\mathcal{Y}$.

Some Background on Bochner Integration In the course of the following statements we use *Bochner integration*. Bochner integration is a straight-forward generalization of Lebesgue integration. Hitherto, we were referring to distributions on finite $\mathcal{Y}$, e.g., $f(x) \in \mathcal{P} \subseteq \mathbb{R}^{|\mathcal{Y}|}$ and computed expectations such as $\mathbb{E}_{D_X}[f(X)]$. Given that $\mathcal{Y}$ is infinite, $f(x)$ is not in a Euclidean space anymore. Hence, we need to pay attention whether $\mathbb{E}_{D_X}[f(X)]$ is well-defined. For that reason, we assume that $\mathcal{P} \subseteq W_{TV}$, where W_{TV} is a Banach space of measures equipped with the total variation distance. By assumption f is measurable and it is easy to see that, $\mathbb{E}_D[\|f(X)\|_{TV}] \leq 1 < \infty$, hence, $\mathbb{E}_{D_X}[f(X)] \in W_{TV}$ is a well-defined Bochner integral (Diestel and Uhl, 1977, II.2 Theorem 2). By definition we have, for all $A \in \mathcal{B}(\mathcal{Y})$,

$$\mathbb{E}_{D_X}[f(X)](A) = \mathbb{E}_{D_X}[f_A(X)],$$

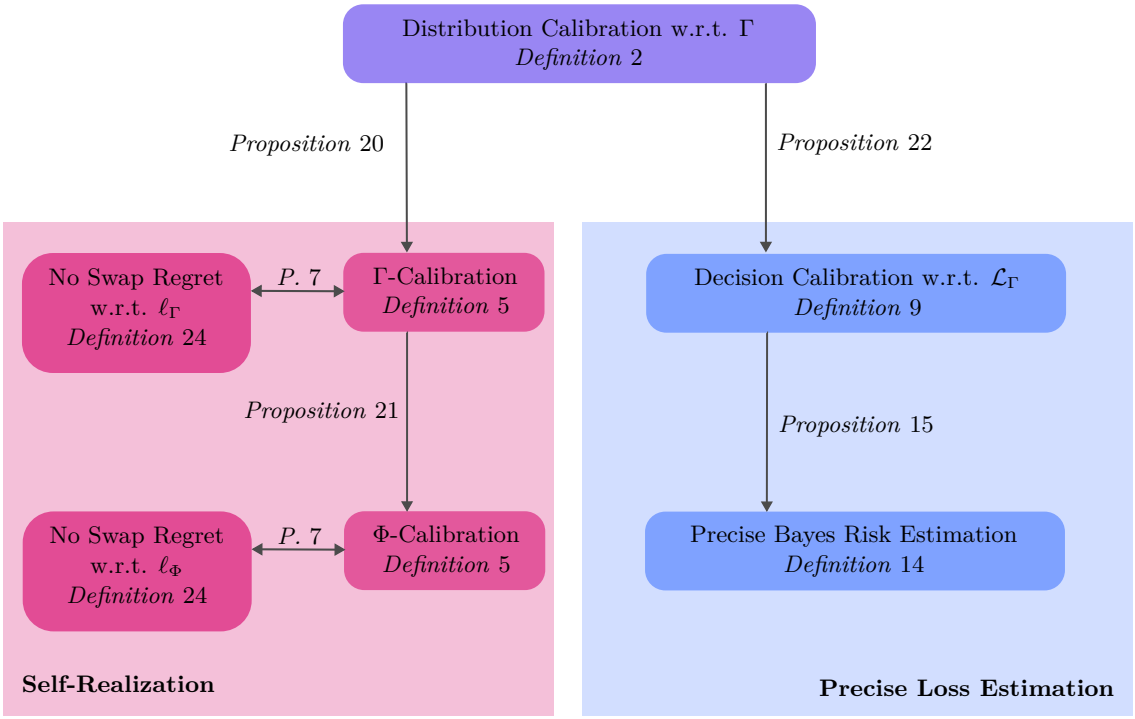


Figure 5: Relationships between Notions of Calibration. Implications under perfect calibration, *infinite* $\mathcal{Y}$ and elicitable property Γ and Φ. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

where $f_A(x)$ denotes the probability assigned to the measurable set A by the probability measure $f(x)$. That is we can rewrite the distribution calibration condition (Equation 4) to

$$D_{Y|\Gamma\circ f(X)=\gamma} = \mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]. \tag{5}$$

Proposition 19 (Distribution Calibration is Inherited (General $\mathcal{Y}$))**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. Suppose, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. If the predictor f is distribution calibrated with respect to Γ on D, then it is distribution calibrated with respect to $\Phi \coloneqq \phi \circ \Gamma$ for all $\phi\colon \mathcal{R} \to \mathcal{R}'$.*

Proof Pick any $v \in \operatorname{supp} D_{\Phi\circ f(X)}$. By definition of the conditional distribution, we have,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{G\sim D_{\Gamma\circ f(X)}}[D_{Y|\Gamma\circ f(X)=G}(A)|\phi(G) = v].$$

As f is distribution calibrated with respect to Γ on D, we have,

$$D_{Y|\Gamma\circ f(X)=G}(A) = \mathbb{E}_{X\sim D_X}[f_A(X)|\Gamma \circ f(X) = G],$$

for all $G \in \operatorname{supp} D_{\Gamma\circ f(X)}$. In particular, that is for all G such that $D_{G\sim\Gamma\circ f(X)}(\phi(G) = v) > 0$, because for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. Hence,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{G\sim D_{\Gamma\circ f(X)}}[\mathbb{E}_{X\sim D_X}[f_A(X)|\Gamma \circ f(X) = G]|\phi(G) = v],$$

Finally,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{X\sim D_X}[f_A(X)|\Phi\circ f(X)=v].$$

■

Proposition 20 (Distribution Calibration with Respect to Γ implies Γ-Calibration (General $\mathcal{Y}$))**.** *Let $\Gamma\colon \mathcal{P}\to\mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X}\times\mathcal{Y}$, regular wrt. $\mathcal{P}$. Suppose that $\mathcal{P}\subseteq W_{TV}$ is a subset of a Banach space equipped with the total variation distance. Let $f\colon \mathcal{X}\to\mathcal{P}$ be a predictor which is distribution calibrated with respect to Γ. Then, $\Gamma\circ f$ is Γ-calibrated.*

Proof We have to show that for all $\gamma\in\operatorname{supp} D_{\Gamma\circ f(X)}$,

$$\Gamma(D_{Y|\Gamma\circ f(X)=\gamma}) = \gamma. \tag{6}$$

Fix any such γ. Based on Corollary 8 in (Diestel and Uhl, 1977) reproduced in (Noarov and Roth, 2023), we have,

$$\mathbb{E}_D[f(X)|\Gamma\circ f(X)=\gamma]\in\overline{\text{co}}f(\{x\in X \text{ s.t. } :\Gamma\circ f(x)=\gamma\}).$$

In particular,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma\circ f(X)=\gamma])\in\Gamma(\overline{\text{co}}f(\{x\in X \text{ s.t. } :\Gamma\circ f(x)=\gamma\}))=\{\gamma\},$$

because Γ has convex level sets. It follows by Equation 5,

$$\Gamma(D_{Y|\Gamma\circ f(X)=\gamma}) = \Gamma(\mathbb{E}_D[f(X)|\Gamma\circ f(X)=\gamma]) = \gamma.$$

■

Proposition 21 (Γ-Calibration is Inherited by Refined Properties (General $\mathcal{Y}$))**.** *Let $\Gamma\colon\mathcal{P}\to\mathcal{R}$ be a property, D a regular data distribution on $\mathcal{X}\times\mathcal{Y}$ with respect to $\mathcal{P}$. Suppose that f is a Γ-predictor which is Γ-calibrated on D. Suppose, furthermore, that for every $x\in\mathcal{X}$, $f(x)\in\operatorname{supp} D_{f(X)}$. Then, for every property Φ with convex level sets and which is refined by Γ, the Φ-predictor $\phi\circ f$, where ϕ is defined following Definition 3, is Φ-calibrated on D.*

Proof We have to show that,

$$\Phi(D_{Y|\phi\circ f(X)=v}) = \phi\circ\Gamma(D_{Y|\phi\circ f(X)=v}) = v,$$

for all $v\in\operatorname{supp} D_{\phi\circ f(X)}$. Fix $v\in\operatorname{supp} D_{\phi\circ f(X)}$ for the following. Let us define the probabilistic predictor $h\colon\mathcal{X}\to\mathcal{P}$ where $x\mapsto D_{Y|f(X)=f(x)}$. The mapping is measurable by definition, as f is measurable by assumption. Let $A\coloneqq\{x\in X\colon\phi\circ f(x)=v\}$. Since, f is Γ-calibrated, and for all $x\in\mathcal{X}$, $f(x)\in\operatorname{supp} D_{f(X)}$, it holds,

$$\Gamma(h(x)) = \Gamma(D_{Y|f(X)=f(x)}) = f(x)$$

for all $x \in A$. Hence, in particular,

$$\Phi(h(x)) = v,$$

for all $x \in A$. By the law of total expectation, we obtain, for all $B \in \mathcal{B}(\mathcal{Y})$,

$$\begin{aligned} D_{Y|\phi\circ f(X)=v}(Y \in B) &= \mathbb{E}_D[[\![Y \in B]\!]|\phi(f(\tilde{X})) = v] \\ &= \mathbb{E}_{G\sim D_{f(X)}}\left[\mathbb{E}_D[[\![Y \in B]\!]|f(X) = G]|\phi(G) = v\right] \\ &= \mathbb{E}_{G\sim D_{f(X)}}\left[\mathcal{D}_{Y|f(X)=G}(Y \in B)|\phi(G) = v\right]. \end{aligned}$$

It follows by definition of h, A and properties of the Bochner integral, $D_{Y|\phi\circ f(X)=v} = \mathbb{E}_{D_X}[h(X)|X \in A]$. Based on Corollary 8 in (Diestel and Uhl, 1977) reproduced in (Noarov and Roth, 2023), we have,

$$\mathbb{E}_D[h(X)|X \in A] \in \overline{\text{co}}h(A).$$

In particular,

$$\Phi(\mathbb{E}_D[h(X)|X \in A]) \in \Phi(\overline{\text{co}}(A) = \{v\},$$

because Φ has convex level sets. ■

Proposition 22 (Distribution Calibration Implies Decision Calibration (General $\mathcal{Y}$))**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor which is distribution calibrated with respect to Γ on D. Then, f is decision calibrated with respect to $\mathcal{L}_\Gamma \coloneqq \{\ell\colon \ell$ is a bounded, $\mathcal{P}$-consistent loss function for $\Gamma\}$.*

Proof Let $\ell \in \mathcal{L}_\Gamma$ be arbitrary. Because ℓ is bounded, the following expectations are well-defined. For every $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$, we have,

$$\begin{aligned} &\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X) = \gamma\right] \\ &= \mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma)|\Gamma\circ f(X) = \gamma\right] - \mathbb{E}_{(X,Y)\sim D}\left[\mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X) = \gamma\right] \\ &= \int_{y\in\mathcal{Y}} \ell(y,\gamma) dD_{Y|\Gamma\circ f(X)=\gamma} - \mathbb{E}_{X\sim D_X}\left[\int_{y\in\mathcal{Y}} \ell(y,\gamma) df(X) \,\middle|\, \Gamma\circ f(X) = \gamma\right] \\ &= 0, \end{aligned}$$

The last equality holds, because, by (Diestel and Uhl, 1977, II.2 Theorem 6) and distribution calibration,

$$\begin{aligned} \mathbb{E}_{X\sim D_X}\left[\int_{y\in\mathcal{Y}} \ell(y,\gamma) df(X) \,\middle|\, \Gamma\circ f(X) = \gamma\right] &= \int_{y\in\mathcal{Y}} \ell(y,\gamma) d\mathbb{E}_{X\sim D_X}\left[f(X) \mid \Gamma\circ f(X) = \gamma\right] \\ &= \int_{y\in\mathcal{Y}} \ell(y,\gamma) dD_{Y|\Gamma\circ f(X)=\gamma}. \end{aligned}$$

It follows, legitimately ignoring $\gamma \in \text{im}\Gamma\circ f$ such that $D_X(\Gamma\circ f(X) = \gamma) = 0$,

$$\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\Gamma(f(X)))]\right] = 0.$$

■

Appendix B. Approximate Calibration

In the above, we left approximate versions of calibration relatively untouched. In the following section, we extend several statements beyond perfect calibration. Figure 6 summarizes the findings of this appendix analogously to Figure 1 for the perfect calibration case.

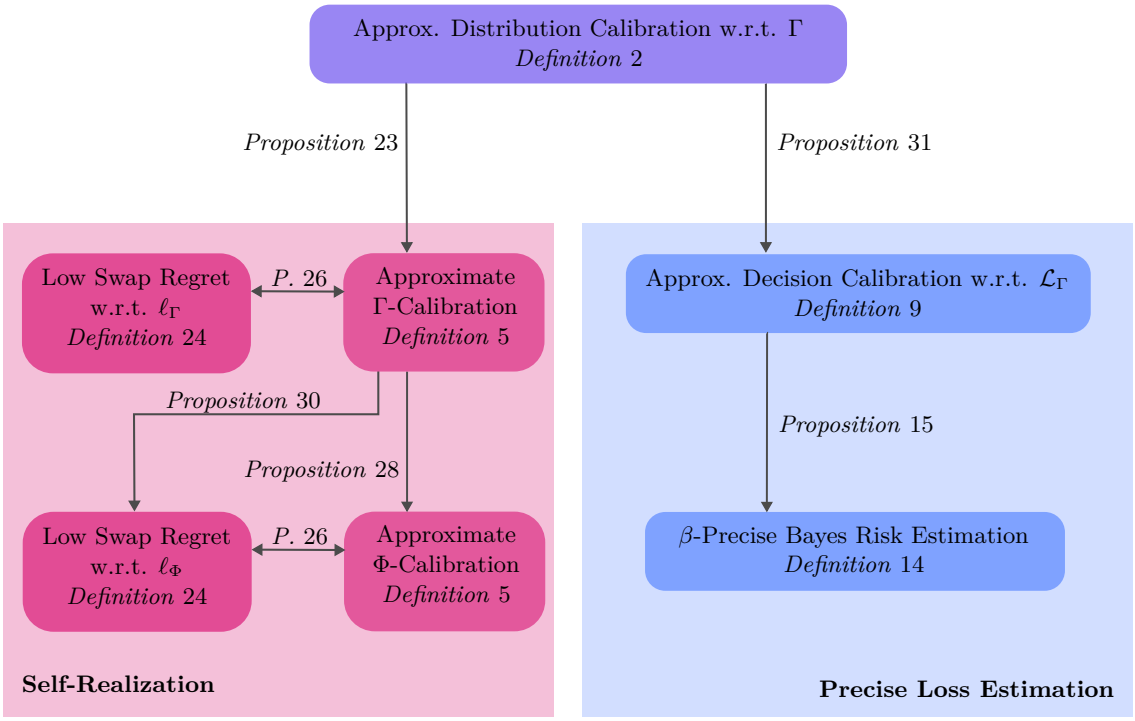


Figure 6: Relationship between Approximate Notions of Calibration. Implications Under Approximate Calibration and Finite $\mathcal{Y}$. Further conditions are stated in the referenced propositions. Those conditions particularly contain Lipschitz and Smoothness assumptions. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

B.1 Approximate Distribution Calibration Implies Approximate Γ-Calibration

In Proposition 6 we have shown that distribution calibration implies property calibration for all properties with convex level sets. We extend this statement to approximate versions of calibration in the following. To this end, we assume that the property Γ is Lipschitz continuous.

Proposition 23 (Approximate Distribution Calibration w.r.t. Γ implies approximate Γ-Calibration for Lipschitz continuous Γ)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ with $\mathcal{R} \subseteq \mathbb{R}$ be a K-Lipschitz property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor that is $\alpha(\gamma)$-approximately distributionally calibrated with respect to Γ. Then, $\Gamma \circ f$ is $|\mathcal{Y}|K\alpha(\gamma)$-approximately Γ-calibrated.*

Proof Since Γ is real-valued, the metric m is the absolute difference. We have to show that for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$|\Gamma(D_{Y|\Gamma \circ f(X)=\gamma}) - \gamma| \leq |\mathcal{Y}| K \alpha(\gamma). \tag{7}$$

Pick any $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$. Lemma 38 applied to D, f and $A \coloneqq \{x \in \mathcal{X} \colon \Gamma \circ f(x) = \gamma\}$ gives,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) = \gamma.$$

Second, the total variation distance between $D_{Y|\Gamma \circ f(X)=\gamma}$ and $\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]$ is bounded above for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$\sup_{A \subseteq \mathcal{Y}} \left| \sum_{y \in A} D_{Y|\Gamma \circ f(X)=\gamma}(Y = y) - \mathbb{E}_D[f_y(X)|\Gamma \circ f(X) = \gamma] \right| \leq |\mathcal{Y}| \alpha(\gamma)$$

where $f_y(X)$ denotes the y-component of the prediction $f(X) \in \Delta(\mathcal{Y})$. This follows from the fact that f is $\alpha(\gamma)$-approximate distribution calibrated with respect to Γ. Hence, by applying Γ and exploiting the Lipschitzness of Γ, we obtain the desired Equation (7). ■

B.2 Approximate Γ-Calibration is Equivalent to Swap Regret

Γ-calibration is equivalent to swap regret not only in the perfect case, but as well in the approximate case. To this end, we formally introduce a distributional definition of swap regret similar in spirit to (Cesa-Bianchi and Lugosi, 2006, p. 91) and introduce additional regularity assumptions on Γ and its identification function V.

Definition 24 (Swap Regret)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be an elicitable property with $\mathcal{P}$-consistent scoring function $\ell \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let $f \colon \mathcal{X} \to \mathcal{R}$ be a Γ-predictor. The Γ-predictor f has $\beta(\gamma)$-bounded* swap regret *on D if for every $\gamma \in \operatorname{supp} D_{f(X)}$,*

$$\mathbb{E}_D\left[\ell(Y, \gamma)|f(X) = \gamma\right] - \min_{\hat{\gamma} \in \operatorname{im}\Gamma} \mathbb{E}_D\left[\ell(Y, \hat{\gamma})|f(X) = \gamma\right] \leq \beta(\gamma).$$

One of the crucial insights in (DeGroot and Fienberg, 1983) was that mean-calibration relates to swap regret for the squared loss function.[22] In this section, we more generally show that if the property, which ought to be calibrated, is identifiable with a certain regular identification function, approximate calibration is equivalent to low swap regret. In particular, this extends the equivalence spelled out in Proposition 7 to the approximate case in Proposition 26, which requires certain regularity conditions on the identification function given in Definition 25.

Definition 25 (Conditions on Identification Functions)**.** *Let $\mathcal{R} \subseteq \mathbb{R}$ be an interval and $\Gamma \colon \mathcal{P} \to \mathcal{R}$ an identifiable property with identification function $V \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Let $\gamma, \gamma' \in \operatorname{im}\Gamma$. The identification function V is called*

22. Globus-Harris et al. (2023) and Gopalan et al. (2024b) extend this relationship to multi-calibration and swap agnostic learners.

oriented *if and only if for all* $P \in P$, $V(P, \gamma) > 0 \Leftrightarrow \gamma > \Gamma(P)$.

locally non-constant *if and only if there exists* $N > 0$ *such that for all* $P \in \mathcal{P}$, $N|\gamma - \Gamma(P)| \leq |V(P, \gamma)|$.

locally Lipschitz *if and only if there exists* $M > 0$ *such that for all* $P \in \mathcal{P}$, $|V(P, \gamma)| \leq M|\gamma - \Gamma(P)|$.

Let us shortly discuss the regularity conditions. Orientedness is arguably a weak assumption. Finocchiaro and Frongillo (2018) have shown that in finite dimensions, i.e., $|\mathcal{Y}| < \infty$, there exists a reweighting of any identification function, such that the reweighted identification function is oriented and identifies the same property.[23] Furthermore, we obtain orientedness for free in the main theorem of (Steinwart et al., 2014). The non-constantness and Lipschitzness assumptions have been considered, in its non-local variant, in the context of composite properties (Noarov and Roth, 2023, Assumption 5.2 and 5.3). Note that $N \leq M$ by definition, which intuitively captures the idea that an identification function cannot be more non-constant than Lipschitz-smooth. In Appendix E we provide several examples of properties beyond the mean which have identification functions which fulfill all of the above properties.

Proposition 26 (Approximate Γ-Calibration Equivalent to Low Swap Regret)**.** *Let* $\Gamma \colon \mathcal{P} \to \mathcal{R}$ *with* $\mathcal{R} \subseteq \mathbb{R}$ *be an identifiable property with oriented, bounded identification function* $V \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ *measurable in both its inputs with respect to the standard Borel-*σ*-algebra on the sets* $\mathcal{Y}, \mathcal{R}$ *and* $\mathbb{R}$*. Then,* Γ *is elicitable with* $\mathcal{P}$*-consistent scoring function* $\ell \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$*,*

$$\ell(y, \gamma) \coloneqq \int_{\gamma_0}^{\gamma} V(y, r) dr + \kappa(y), \tag{8}$$

for some $\gamma_0 \in \mathrm{im}\Gamma$ *and* $\kappa \colon \mathcal{Y} \to \mathbb{R}$ *having a finite expected value with respect to all* $P \in \mathcal{P}$*. Let* f *be a* Γ*-predictor and* D *a regular data distribution.*

1. *Suppose* V *is locally Lipschitz on* $\mathcal{P}$ *with parameter* M*. If* f *is* $\alpha(\gamma)$*-approximately* Γ*-calibrated, then* f *has* $\beta(\gamma) \coloneqq \frac{M}{2}\alpha(\gamma)^2$*-bounded swap regret.*
2. *Suppose* V *is locally non-constant on* $\mathcal{P}$ *with parameter* N*. If* f *has* $\beta(\gamma)$*-bounded swap regret, then* f *is* $\alpha(\gamma) \coloneqq \sqrt{\frac{2}{N}\beta(\gamma)}$*-approximate calibrated.*

Proof Lemma 27 does almost all of the heavy lifting. It remains to apply Equation 10 to $D_{Y|f(X)=\gamma}$ for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$|\ell(D_{Y|f(X)=\gamma}, \gamma) - \ell(D_{Y|f(X)=\gamma}, \Gamma(D_{Y|f(X)=\gamma}))| \leq \frac{M}{2}\left(\gamma - \Gamma(D_{Y|f(X)=\gamma})\right)^2 \leq \frac{M}{2}\alpha(\gamma)^2,$$

to obtain the first statement. For the second statement, we rewrite Equation 11 for $D_{Y|f(X)=\gamma}$ for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$|\gamma - \Gamma(D_{Y|f(X)=\gamma})| \leq \sqrt{\frac{2}{N}|\ell(D_{Y|f(X)=\gamma}, \gamma) - \ell(D_{Y|f(X)=\gamma}, \Gamma(D_{Y|f(X)=\gamma}))|} \leq \sqrt{\frac{2}{N}\beta(\gamma)}.$$

23. Finocchiaro and Frongillo (2018) show that the reweighted identification function is monotone increasing, which implies orientedness.

■

Lemma 27 (Identification Function Defines Consistent Loss Functions)**.** *Let* $\Gamma\colon \mathcal{P} \to \mathcal{R}$ *with* $\mathcal{R} \subseteq \mathbb{R}$ *be an identifiable property with oriented, bounded identification function* $V\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ *as in Proposition 26. Then,* Γ *is elicitable with* $\mathcal{P}$*-consistent scoring function* $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$*,*

$$\ell(y, \gamma) \coloneqq \int_{\gamma_0}^{\gamma} V(y, r)dr + \kappa(y), \tag{9}$$

for some $\gamma_0 \in \mathrm{im}\Gamma$ *and* $\kappa\colon \mathcal{Y} \to \mathbb{R}$ *having a finite expected value with respect to all* $P \in \mathcal{P}$*.*

(a) If furthermore, V *is locally Lipschitz with parameter* M*, then* ℓ *is locally Hölder-smooth with parameters* $(\frac{M}{2}, 2)$*, i.e., for all* $\gamma \in \mathcal{R}$*,*

$$|\ell(P, \gamma) - \ell(P, \Gamma(P))| \leq \frac{M}{2} \left(\gamma - \Gamma(P)\right)^2. \tag{10}$$

(b) If furthermore, V *is locally non-constant with parameter* N*, then* ℓ *is locally anti Hölder-smooth with parameters* $(\frac{N}{2}, 2)$*, i.e., for all* $\gamma \in \mathcal{R}$*,*

$$\frac{N}{2} \left(\gamma - \Gamma(P)\right)^2 \leq |\ell(P, \gamma) - \ell(P, \Gamma(P))|. \tag{11}$$

Proof The first statement has been proved in (Steinwart et al., 2014). We provide a reiteration of the argument for the sake of completeness, and because we will reuse some equations for the additional statements.

For any $P \in \mathcal{P}$, we want to show that $\mathbb{E}_{Y\sim P}[\ell(Y, \gamma)] > \mathbb{E}_{Y\sim P}[\ell(Y, \Gamma(P))]$ for all $\gamma \in \mathcal{R}$, $\gamma \neq \Gamma(P)$. Let us first consider the case that $\gamma > \Gamma(P)$. Hence,

$$\begin{aligned}
&\mathbb{E}_{Y\sim P}[\ell(Y, \gamma)] - \mathbb{E}_{Y\sim P}[\ell(Y, \Gamma(P))] \\
&= \mathbb{E}_{Y\sim P}[\ell(Y, \gamma) - \ell(Y, \Gamma(P))] \\
&= \mathbb{E}_{Y\sim P}\left[\int_{\Gamma(P)}^{\gamma} V(y, r)dr\right].
\end{aligned}$$

Since V is measurable and bounded in both variables we can apply Fubini's theorem, this gives

$$\mathbb{E}_{Y\sim P}\left[\int_{\Gamma(P)}^{\gamma} V(y, r)dr\right] = \int_{\Gamma(P)}^{\gamma} \mathbb{E}_{Y\sim P}[V(y, r)]dr, \tag{12}$$

finally, since V is oriented,

$$\int_{\Gamma(P)}^{\gamma} \mathbb{E}_{Y\sim P}[V(y, r)]dr = \int_{\Gamma(P)}^{\gamma} V(P, r)dr > 0.$$

The case $\gamma < \Gamma(P)$ follows analogously.

(a) Let $\gamma > \Gamma(P)$, then,

$$\begin{aligned}|\ell(P,\gamma) - \ell(P,\Gamma(P))| &= \left|\int_{\Gamma(P)}^{\gamma} V(P,r)dr\right| \\ &= \int_{\Gamma(P)}^{\gamma} |V(P,r)|dr \\ &\leq \int_{\Gamma(P)}^{\gamma} M|r-\Gamma(P)|dr \\ &= \frac{M}{2}(\gamma-\Gamma(P))^2.\end{aligned}$$

The analogous computation with swapped signs holds for $\gamma < \Gamma(P)$.

(b) Let $\gamma > \Gamma(P)$, then,

$$\begin{aligned}|\ell(P,\gamma) - \ell(P,\Gamma(P))| &= \left|\int_{\Gamma(P)}^{\gamma} V(P,r)dr\right| \\ &= \int_{\Gamma(P)}^{\gamma} |V(P,r)|dr \\ &\geq \int_{\Gamma(P)}^{\gamma} N|r-\Gamma(P)|dr \\ &= \frac{N}{2}(\gamma-\Gamma(P))^2.\end{aligned}$$

The analogous computation with swapped signs holds for $\gamma < \Gamma(P)$.

■

Hence, for the right choice of loss function, Γ-calibration and low swap regret can be used equivalently. Along the lines of (Gopalan et al., 2024b, Theorem 3.3), we extend the equivalence relationship to group-wise definitions of calibration via (Noarov and Roth, 2023, Definition 2.10) in Appendix F. However, we do *not* achieve a direct generalization of (Gopalan et al., 2024b, Theorem 3.3 (1.) swap multicalibration $\iff$ (3.) swap-agnostic learner) by going from the mean (as in their work) to more general identifiable properties. Hence, we provide a bridge from calibration for general identifiable properties to swap regret notions, going beyond (Noarov and Roth, 2023). Nevertheless, the mean is a good example to illustrate Proposition 26.

Example 3. *Let Γ be the mean and $V(y,\gamma) \coloneqq \gamma - y$ its identification function. In this case, $N = M = 1$. The mean is elicited by the squared loss,*

$$\ell(y,\gamma) = \frac{1}{2}(y-\gamma)^2 = \int_0^{\gamma} V(y,r)dr + \frac{1}{2}y^2.$$

Hence, a Γ-predictor f on a regular data distribution D has $\frac{1}{2}\alpha(\gamma)^2$ swap regret if, and only if it is $\alpha(\gamma)$-approximately calibrated (cf. (Gopalan et al., 2024b, Theorem 3.3)).

B.3 Inheritance of Approximate Γ-Calibration

For the approximate analogue of Proposition 8 we require the refined property to be Lipschitz.

Proposition 28 (Approximate Γ-Calibration is Inherited by Refined Properties)**.** *Let Γ be a property, D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, f a Γ-predictor and $\Phi = \phi \circ \Gamma$ a property refined by Γ with convex level sets. Assume further that ϕ is Lipschitz (with constant K), $\mathcal{Y}$ is finite and for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. If f is $\alpha(\gamma)$-approximately Γ-calibrated on D, then $\phi \circ f$ is $C\alpha'(\phi(v))$-approximately Φ-calibrated, where $\alpha'(v) := \sup_{\gamma \in \phi^{-1}(v)}$.*

Proof It is given that,

$$|\Gamma(D_{Y|f(X)=\gamma}) - \gamma| \leq \alpha(\gamma),$$

from which simply follows,

$$|\phi \circ \Gamma(D_{Y|f(X)=\gamma}) - \phi(\gamma)| \leq K\alpha(\gamma).$$

Then, by Proposition 8,

$$|\phi \circ \Gamma(D_{Y|\phi \circ f(X)=v}) - v| \leq K\alpha'(v),$$

where $\alpha'(v) = \sup_{\gamma \in \phi^{-1}(v)} \alpha(\gamma)$. ■

The Lipschitz-condition on the property is strong. For instance, it rules out discrete properties. However, we will argue that in the case of elicitable properties we can still give guarantees about self-realization in terms of low swap regret (Proposition 30). This statement requires an arguably mild regularity condition on the loss function.

Definition 29 (Locally Outcome Lipschitz Loss Function)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$. Suppose a loss function $\ell \colon \mathcal{Y} \times \mathcal{R}' \to \mathbb{R}$ elicits a property $\Phi \colon \mathcal{P} \to \mathcal{R}'$. It is called* locally outcome Lipschitz with respect to Γ *with constant B, if for all $v \in \operatorname{im}\Phi$, $\gamma \in \operatorname{im}\Gamma$, $P \in \Gamma^{-1}(\gamma)$ and $P' \in \mathcal{P}$,*

$$|\ell(P, v) - \ell(P', v)| \leq Bm(\Gamma(P'), \gamma). \tag{13}$$

If $\mathcal{Y}$ is finite, Γ is the full distribution, i.e., $\mathcal{R} = \Delta(\mathcal{Y})$ and ℓ a bounded function, then ℓ is locally outcome Lipschitz (Appendix, Lemma 39). Furthermore, the assumptions made in (Gopalan et al., 2024b) imply Definition 29 for Γ being the mean on a binary outcome set.

Proposition 30 (Low Swap Regret Guarantee for Refined Properties)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property, D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$ and f a Γ-predictor which is Γ-calibrated on D. Let $\Phi \colon \mathcal{P} \to \mathcal{R}'$ be a property which is refined by Γ, i.e., $\Phi \coloneqq \phi \circ \Gamma$ for some $\phi : \mathcal{R} \to \mathcal{R}'$.*

If Φ is elicitable with loss function ℓ, locally outcome Lipschitz with respect to Γ with constant B, and if f is $\alpha(\gamma)$-approximately Γ-calibrated, then the swap regret of $\phi \circ f$ with respect to ℓ is bounded above by $2B\mathbb{E}_{\gamma \sim D_{f(X)}}[\alpha(\gamma) | \phi \circ \gamma = v]$ for all $v \in \operatorname{supp} D_{\phi \circ f(X)}$.

Proof Let $P \in \Gamma^{-1}(\gamma)$, by Definition 29 Equation (13), for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$\begin{aligned}
&\mathbb{E}_D\left[\ell(Y, \phi(\gamma)) - \ell(Y, \Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})|f(X) = \gamma\right] \\
&= \ell(D_{Y|f(X)=\gamma}, \phi(\gamma)) - \ell(D_{Y|f(X)=\gamma}, \Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})) \\
&= \ell(D_{Y|f(X)=\gamma}, v) - \ell(P, \Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})) \\
&\quad + \ell(P, \Phi(D_{Y|\phi\circ f(X)=v})) - \ell(D_{Y|f(X)=\gamma}, \Phi(D_{Y|\phi\circ f(X)=v})) \\
&\leq \ell(D_{Y|f(X)=\gamma}, v) - \ell(P, \Phi(D_{Y|\phi\circ f(X)=v})) + Bm(\Gamma(D_{Y|f(X)=\gamma}), \gamma)| \\
&\leq \ell(D_{Y|f(X)=\gamma}, v) - \ell(P, v) + \ell(P, v) - \ell(P, \Phi(D_{Y|\phi\circ f(X)=v})) \\
&\quad + Bm(\Gamma(D_{Y|f(X)=\gamma}), \gamma) \\
&\leq \ell(P, v) - \ell(P, \Phi(D_{Y|\phi\circ f(X)=v})) + 2Bm(\Gamma(D_{Y|f(X)=\gamma}), \gamma) \\
&\leq 2B\alpha(\gamma).
\end{aligned}$$

It follows that, for all $v \in \operatorname{supp} D_{\phi\circ f(X)}$,

$$\begin{aligned}
&\mathbb{E}_D\left[\ell(Y, \phi(\gamma)) - \ell(Y, \Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})|\phi \circ f(X) = v\right] \\
&\leq 2B\mathbb{E}_{\gamma\sim D_{f(X)}}\left[\alpha(\gamma)|\phi \circ f(X) = v\right].
\end{aligned}$$

■

An analogous proof holds for distributional calibration with respect to Γ and a locally outcome Lipschitz loss function with respect to Γ. The proof is an adapted version of parts of the proof of the main theorem in (Gopalan et al., 2024b).

The inheritance of self-realization is closely connected to "(swap) omniprediction" introduced by (Gopalan et al., 2022b, 2024b). Informally, a (swap) omnipredictor provides predictions which allow for a simple post-processing to achieve low regret against a comparison base line for a large variety of losses. In particular, (swap) omnipredictors involve "for all" quantifiers about a set of losses.

Non-Refining (Swap) Omniprediction Proposition 26 already involves an "omni"-type statement. The statement holds for all loss functions defined following Equation (8). Yet, there is no property refinement necessary.

Refining (Swap) Omniprediction The inheritance rule makes use of the refinement function ϕ. This refinement function generalizes the post-processing function k_ℓ in (Gopalan et al., 2022b). In particular, Proposition 30 can be plugged together for a set of loss functions. For instance, let f be a Γ-calibrated Γ-predictor on a data distribution D. Then, those predictions guarantee low swap regret for all loss functions which are locally outcome Lipschitz with respect to Γ. Hence, this is an "omni"-type regret statement, related to the main theorem in (Gopalan et al., 2024b). In there, the authors focused on a very specific set of losses which are convex, Lipschitz and locally outcome Lipschitz with respect to the mean. However, their main theorem is generalized to multi-calibration with respect to groups. We push aside this further complexity in this work (cf. Section 7). Hence, our statement is less general in the scope of groupings, but more general in the scope of sets of loss functions and predicted properties.[24]

24. Note that (Lu et al., 2025) as well consider generalizations of the set of loss functions.

B.4 Approximate Distribution Calibration Implies Approximate Decision Calibration

Subject to mild assumptions on the boundedness of the loss function ℓ which elicits the property of interest Γ, Proposition 12 can be generalized from perfect to approximate calibration.

Proposition 31 (Approximate Distribution Calibration Implies Approximate Decision Calibration)**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$ Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor $\alpha(\gamma)$-approximately distribution calibrated with respect to Γ on D. Then, f is $C\mathbb{E}_{\gamma\sim D_{\Gamma\circ f(X)}}[\alpha(\gamma)]$-approximate decision calibrated with respect to $\mathcal{L}_{\Gamma,C} \coloneqq \{\ell\colon \ell$ is $\mathcal{P}$-consistent loss function for Γ and $\sup_{r\in\mathcal{R}} \sum_{y\in\mathcal{Y}} |\ell(y,r)| \leq C < \infty\}$.*

Proof The predictor f is $\alpha(\gamma)$-*approximate distribution calibrated with respect to* Γ on D if for every $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$,

$$\left| D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_D[f_y(X)|\Gamma\circ f(X)=\gamma] \right| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y},$$

where $f_y(x) \in [0,1]$ denotes the y-component of the prediction $f(x) \in \Delta(\mathcal{Y})$ for $x \in \mathcal{X}$.

Let $\ell \in \mathcal{L}_{\Gamma,C}$ be arbitrary but fixed. For all $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$, we have,

$$\begin{aligned}
&\left|\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X)=\gamma\right]\right| \\
&= \left|\sum_{y\in\mathcal{Y}} \left(D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma]\right)\ell(y,\gamma)\right| \\
&\leq \sum_{y\in\mathcal{Y}} \left|D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma]\right| |\ell(y,\gamma)| \\
&\leq \alpha(\gamma) \sum_{y\in\mathcal{Y}} |\ell(y,\gamma)| \leq C\alpha(\gamma).
\end{aligned}$$

The result follows by taking the expectation over $\gamma \sim D_{\Gamma\circ f(X)}$. For detailed steps from the first to the second term see the proof of Proposition 12. ■

Appendix C. Distribution Calibration with Respect to all Binary Properties Implies Full Distribution Calibration

It follows relatively immediately that distribution calibration with respect to the full distribution implies distribution calibration with respect to any property. The reverse direction requires one to think about a *set* of properties. We show that if a forecast is perfectly distribution calibrated with respect to all binary, elicitable properties, it is distribution calibrated with respect to the full distribution. This statement requires some technical-looking assumption on separability via hyperplanes. If the number of different predicted forecasts is finite, which is arguably a mild assumption in practice, then this separability assumption holds.

Lemma 32. *If $|\mathrm{im}f| < \infty$, then for every $p \in \mathrm{im}f$ there exists a hyperplane $H = \{x \in \mathcal{R}^{|\mathcal{Y}|} \colon \langle a, x\rangle = b\}$ such that $p \in H$ and $0 < \epsilon \leq \inf_{h\in H} \|h - q\|$ for every $q \in \mathrm{im}f \setminus \{p\}$.*

Proof The proof follows by contradiction. Let us assume that for some point $p \in \mathrm{im}f$ there is no hyperplane such that $p \in H$ and $0 < \epsilon \leq \inf_{h\in H} \|h - q\|$ for every $q \in \mathrm{im}f \setminus \{p\}$. That is, for every hyperplane H such that $p \in H$ there is $q_H \in \mathrm{im}f \setminus \{p\}$ with $\inf_{h\in H} \|h - q\| < \epsilon$ for every $\epsilon > 0$. It follows that such $q_H \in H$, because $\inf_{h\in H} \|h - q\| = 0$. Now, since there are uncountably many such hyperplanes H but $|\mathrm{im}f| < \infty$ we get into a contradiction. ■

Proposition 33 (Recovering Distribution Calibration by Distribution Calibration with respect to all Binary, Elicitable Properties)**.** *Let D be a regular data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite. Let $f \colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. Suppose that $p_D(f(X) = q) > 0$ for all $q \in \mathrm{im}f$.*

If the predictor f is distribution calibrated with respect to all binary elicitable properties $\Phi \colon \Delta(\mathcal{Y}) \to \{0, 1\}$, then it is distribution calibrated with respect to the full distribution on D.

Proof Let us pick an arbitrary $p \in \mathrm{im}f$. We show that there exist two elicitable, binary properties $\underline{\Phi}_p, \overline{\Phi}_p$ such that,

$$\begin{aligned} &\underline{\Phi}_p(q) = \overline{\Phi}_p(q), \quad \forall q \in \mathrm{im}f \setminus \{p\} \\ &\underline{\Phi}_p(p) \neq \overline{\Phi}_p(p). \end{aligned}$$

Then, we can argue that there is $a \in \{0, 1\}$ such that

$$\begin{aligned} \mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \underline{\Phi}_p \circ f(X) = a] = 0 \\ \mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \overline{\Phi}_p \circ f(X) = a] = 0, \end{aligned}$$

where, with some $c \in [0, 1]$,

$$\begin{aligned} &\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \overline{\Phi}_p \circ f(X) = a] \\ &= c\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \underline{\Phi}_p \circ f(X) = a] + p_D(f(X) = p)\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | f(X) = p]. \end{aligned}$$

Hence, because $p_D(f(X) = p) > 0$,

$$\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | f(X) = p] = 0.$$

Since $p \in \mathrm{im}f$ was picked arbitrarily, this shows the claim.

Nevertheless, it remains to argue that there exist such two elicitable, binary properties $\underline{\Phi}_p, \overline{\Phi}_p$ as demanded above. We show this by the prooving the existence of two hyperplanes such that all points $q \in \mathrm{im}f$ except for p stay on the same side of both hyperplanes, but p switches the side. As proven in (Lambert, 2019, Theorem 1), properties are elicitable if and only if the level sets form a power diagram, i.e., the intersection of the simplex $\Delta(\mathcal{Y}) \subseteq \mathbb{R}^{|\mathcal{Y}|}$ with a Voronoi diagram. For our purpose, it suffices to know that a hyperplane in $\mathbb{R}^{|\mathcal{Y}|}$ defines a binary, elicitable property. In fact, every binary elicitable property can be expressed through a corresponding hyperplane. Let

$$\underline{\Phi}_p(q) = \begin{cases} 0 \text{ if } \langle a, q\rangle > b - \frac{\epsilon\|a\|}{2} \\ 1 \text{ otherwise.} \end{cases} .$$

and

$$\overline{\Phi}_p(q) = \begin{cases} 0 \text{ if } \langle a, q\rangle > b + \frac{\epsilon\|a\|}{2} \\ 1 \text{ otherwise.} \end{cases} .$$

Then,

$$\underline{\Phi}_p(p) = 0 \neq 1 = \overline{\Phi}_p(p),$$

but for all $q \in \text{im} f \setminus \{p\}$,

$$\underline{\Phi}_p(p) = \overline{\Phi}_p(p),$$

because if $\underline{\Phi}_p(q) = 0$ (the analogous argument holds for $\underline{\Phi}_p(q) = 1$), then

$$\langle a, q\rangle > b - \frac{\epsilon\|a\|}{2},$$

which implies,

$$\langle a, q\rangle \geq b + \|a\|\epsilon > b + \frac{\epsilon\|a\|}{2}$$

by formula of distance to a hyperplane,

$$\epsilon \leq \inf_{h \in H} \|h - q\| = \frac{|\langle a, q\rangle - b|}{\|a\|} \qquad \Leftrightarrow \|a\|\epsilon + b \leq \langle a, q\rangle.$$

■

Appendix D. A Note on Sensibility for Calibration

The reader familiar with (Noarov and Roth, 2023) might question the use of a general property Γ in Definition 5. The authors introduce "sensibility" for calibration which requires that the true property predictor is calibrated with respect to Definition 5. As the authors show therein, a property is only sensible for calibration if and only if the level sets (pre-images) of property values are convex. Steinwart et al. (2014) in turn, have proven that if the property is strictly locally non-constant, a minor technical assumption, and continuous, then convexity of the level sets of the property is equivalent to elicitability (respectively identifiability). If the property is discrete and its image finite, and the outcome set is finite, then convex level sets are not sufficient for the elicitability of Γ (Lambert, 2019, Pg. 12).[25] Hence, even if we require Γ to be sensible to calibration, Definition 5 extends to discrete properties with convex level sets, which are potentially neither identifiable nor elicitable. This justifies the definition of Γ-calibration with respect to any Γ with convex level sets. Furthermore, even if a property Γ is not sensible for calibration, a Γ-predictor could still be Γ-calibrated. However, in this case optimal individual prediction via the true property predictor and calibration are *not* compatible. Nevertheless, Γ-calibration is still a fulfillable criterion.

25. Instead the level sets have to form a Voronoi diagram (Lambert, 2019, Theorem 1).

Appendix E. Examples of Properties with Regular Identification Functions

A priori it is unclear whether there exist interesting, non-trivial, identifiable properties with identification functions fulfilling the assumptions put forward in Definition 25. For this reason we provide a short list of examples.

Example 4. Mean *An easy example for a property with a monotone, locally non-constant, locally Lipschitz identification function is the mean, with identification function* $V(y, \gamma) \coloneqq (y - \gamma)$*, because*

$$\mathbb{E}_{Y\sim P}[(Y - \gamma)] = \Gamma(P) - \gamma.$$

Quantiles *One identification function for a* τ*-quantile is* $V_\tau(y, \gamma) \coloneqq (1-\tau)[\![y < \gamma]\!] - \tau[\![y > \gamma]\!]$*. For a loss function, which elicits the* τ*-quantile see e.g., (Rockafellar and Uryasev, 2002). Assuming that* P *is atomless,*

$$\begin{aligned} V(P, \gamma) &= \mathbb{E}_{Y\sim P}[(1-\tau)[\![Y < \gamma]\!] - \tau[\![Y > \gamma]\!]] \\ &= (1-\tau)\mathbb{E}_{Y\sim P}[[\![Y < \gamma]\!]] - \tau\mathbb{E}_{Y\sim P}[[\![Y > \gamma]\!]] \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y \leq \gamma)) \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y < \gamma) - P(Y = \gamma)) \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y < \gamma)) \\ &= P(Y < \gamma) - \tau. \end{aligned}$$

Hence, the identification function is monotone. For local Lipschitzness we have to assume that the cumulative density function of P *is Lipschitz with parameter* L *around the* τ*-quantile of* P*. For instance, this is true for* β*-distributions. Then, we get*

$$|P(Y < \gamma) - \tau| = |P(Y < \gamma) - P(Y < \gamma^*)| \leq L|\gamma - \gamma^*|,$$

where γ^* *is the* τ*-quantile. Finally, if the cumulative distribution function induced through* P *is locally non-constant with parameter* $N > 0$ *around the* τ*-quantile of* P*, then via the same argument it follows that the identification function is locally non-constant (e.g.,* β*-distribution). In summary, the identification function* V_τ *for the* τ*-quantile inherits the continuity properties from the cumulative distribution function of the probability distribution. Hence, the assumption is fulfilled when restricting the space of possible conditional probability distributions defined through the regular data distribution* D*.*

Ratios of Expectations *Let* $g, h \colon \mathcal{Y} \to \mathbb{R}$ *be measurable functions and* $N < h(y) < M$ *for all* $y \in \mathcal{Y}$*. The ratio of expectations* $\Gamma(P) = \frac{\mathbb{E}_P g(Y)}{\mathbb{E}_P h(Y)}$ *is identified by,*

$$V(y, \gamma) = h(y)\gamma - g(y),$$

because

$$V(P, \gamma) = \mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)] = 0$$

if and only if,

$$\frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]} = \gamma.$$

Furthermore, V is oriented, locally Lipschitz with M,

$$\begin{aligned}
|V(P,\gamma)| &= |\mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)]| \\
&= \left|\mathbb{E}_P[h(Y)]\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&= |\mathbb{E}_P[h(Y)]|\left|\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&\leq M\,|\gamma - \Gamma(P)|\,.
\end{aligned}$$

and locally non-constant with N,

$$\begin{aligned}
|V(P,\gamma)| &= |\mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)]| \\
&= \left|\mathbb{E}_P[h(Y)]\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&= |\mathbb{E}_P[h(Y)]|\left|\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&\geq N\,|\gamma - \Gamma(P)|\,.
\end{aligned}$$

Mean and Variance *The identification function of the mean M is $V(y,\gamma) \coloneqq (y-\gamma)$. On the v-level set of the mean the variance $\mathbb{V}$ can be identified with $V_v(y,\gamma) \coloneqq y^2 - v^2 - \gamma$. Note that for $P \in M^{-1}(v)$,*

$$V_v(P,\gamma) = \mathbb{E}_{Y\sim P}[Y^2 - v^2 - \gamma] = \mathbb{V}(P) - \gamma.$$

Quantile and CVar *The identification function of a τ-quantile Q_τ is given above. The CVar_τ is identified on the v-level set of the τ-quantile with $V_v(y,\gamma) = v + \frac{1}{1-\tau}\max(0, y - v) - \gamma$. Note that for $P \in Q_\tau^{-1}(v)$*

$$\begin{aligned}
V_v(P,\gamma) &= \mathbb{E}_{Y\sim P}\left[v + \frac{1}{1-\tau}\max(0, Y - v)\right] \\
&= v + \frac{1}{1-\tau}\mathbb{E}_{Y\sim P}[\max(0, Y - v)] - \gamma \\
&= \mathrm{CVar}_\tau(P) - \gamma,
\end{aligned}$$

The second last line follows from (Rockafellar and Uryasev, 2002, Theorem 10).

Appendix F. Calibration-Swap Regret Bridge under Groups

We call $c\colon \mathcal{X} \to \{0,1\}$ a *group* if c is measurable and $D(c(X)=1) > 0$. The set $\mathcal{C}$ denotes a collection of groups.

Definition 34 ($\mathcal{C}$-Robust Γ-Swap-Learner)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation* (8)*. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. A Γ-predictor f is a β-approximate $\mathcal{C}$-robust Γ-swap-learner on D, iff, for all $c \in \mathcal{C}$,*

$$\mathbb{E}_D\left[\ell(Y, f(X))|c(X) = 1\right] \leq \min_{h \in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_D\left[\ell(Y, h(X))|c(X) = 1\right] + \frac{\beta}{\mathbb{E}_D\left[c(X)\right]},$$

where $\mathcal{M}_f(\mathcal{X}, \mathcal{R})$ is the set of all measurable functions from $(\mathcal{X}, \sigma(f))$ to $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$, where $\sigma(f)$ is the σ-algebra induced by the Γ-predictor f and $\mathcal{B}(\mathcal{R})$ the Borel-σ-algebra on $\mathcal{R}$.

The definition of a $\mathcal{C}$-robust swap-learner inspired by minimax regret in distributional robust optimization. Formalized as an optimization problem the regret here is a special case of a minimax regret optimization problem as in (Agarwal and Zhang, 2022). The class of functions $\mathcal{F}$ in their Equation (1) is $\mathcal{M}_f(\mathcal{X}, \mathcal{R})$ in our case and the set of probability distributions $\mathcal{P}$ in their Equation (1) corresponds to the conditional probabilities on $\mathcal{X} \times \mathcal{Y}$ induced by conditioning on the groups $\mathcal{C}$ in our case. The predictor $f \in \mathcal{M}_f(\mathcal{X}, \mathcal{R})$.

Proposition 35 (l^2-Multicalibration implies Robust Swap-Learner)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation* (8)*. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be α-approximately $(\mathcal{C}, \nu)$-multicalibrated following Definition 15 in (Noarov and Roth, 2023). Then, f is a $\frac{M}{2}\alpha$-approximate $\mathcal{C}$-robust Γ-swap-learner.*

Proof

$$\begin{aligned}
\alpha &\geq \sup_{c \in C} \mathbb{E}_{X \sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}}[|\Gamma(D_{Y|f(X)=\gamma, c(X)=1}) - \gamma|^2] \\
&\overset{P26}{\geq} \frac{2}{M} \sup_{c \in C} \mathbb{E}_{X \sim D_X}[c(X)] \cdot \\
&\mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}} \left[\mathbb{E}_D\left[\ell(Y, \gamma)|f(X) = \gamma, c(X) = 1\right] - \min_{\hat{\gamma} \in \mathrm{im}\Gamma} \mathbb{E}_D\left[\ell(Y, \hat{\gamma})|f(X) = \gamma, c(X) = 1\right]\right] \\
&= \frac{2}{M} \sup_{c \in C} \mathbb{E}_D\left[c(X)\ell(Y, f(X))\right] \\
&\quad - \mathbb{E}_{X \sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}}[\min_{\hat{\gamma} \in \mathrm{im}\Gamma} \mathbb{E}_D\left[\ell(Y, \hat{\gamma})|f(X) = \gamma, c(X) = 1\right]] \\
&= \frac{2}{M} \sup_{c \in C} \mathbb{E}_D\left[c(X)\ell(Y, f(X))\right] - \min_{h \in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_D\left[c(X)\ell(Y, h(X))\right].
\end{aligned}$$

The final step is a consequence of conditional expectations and the following argument. For every $c \in \mathcal{C}$, applying (Rockafellar and Wets, 2009, Theorem 14.60) with $T = \{x \in$

$\mathcal{X}\colon c(x)=1\}$ and $\mathcal{A}=\mathcal{B}(\mathcal{X})\cap T$ gives,

$$\begin{aligned}
&\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\min_{\hat{\gamma}\in \mathrm{im}\Gamma}\mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma,c(X)=1\right]\right]\\
&\qquad=\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\min_{\alpha\in\mathbb{R}}\mathbb{E}_D\left[\ell(Y,\alpha)|f(X)=\gamma,c(X)=1\right]\right]\\
&\qquad=\min_{h\in\mathcal{M}_f(\mathcal{X},\mathcal{R})}\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\mathbb{E}_D\left[\ell(Y,h(X))|f(X)=\gamma,c(X)=1\right]\right]\\
&\qquad=\min_{h\in\mathcal{M}_f(\mathcal{X},\mathcal{R})}\mathbb{E}_D\left[\ell(Y,h(X))|c(X)=1\right].
\end{aligned}$$

■

Proposition 36 (Robust Swap-Learner implies l^2-Multicalibration)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation* (8)*. Let D be a data distribution on $\mathcal{X}\times\mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be a β-approximate $\mathcal{C}$-robust Γ-swap-learner. Then, f is $\frac{2L^2}{N}\beta$-approximately $(\mathcal{C},\nu)$-multicalibrated following Definition 15 in (Noarov and Roth, 2023).*

Proof

$$\begin{aligned}
\beta &\geq \sup_{c\in C}\mathbb{E}_D\left[c(X)\ell(Y,f(X))\right]-\min_{h\in\mathcal{M}_f(\mathcal{X},\mathcal{R})}\mathbb{E}_D\left[c(X)\ell(Y,h(X))\right]\\
&=\sup_{c\in C}\mathbb{E}_D\left[c(X)\ell(Y,f(X))\right]\\
&\quad-\mathbb{E}_{X\sim D_X}[c(X)]\cdot\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\min_{\hat{\gamma}\in\mathrm{im}\Gamma}\mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma,c(X)=1\right]\right]\\
&=\sup_{c\in C}\mathbb{E}_{X\sim D_X}[c(X)]\cdot\\
&\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\mathbb{E}_D\left[\ell(Y,\gamma)|f(X)=\gamma,c(X)=1\right]-\min_{\hat{\gamma}\in\mathrm{im}\Gamma}\mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma,c(X)=1\right]\right]\\
&\overset{P26}{\geq}\frac{N}{2}\sup_{c\in C}\mathbb{E}_{X\sim D_X}[c(X)]\cdot\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}\left[\left|\Gamma(D_{Y|f(X)=\gamma,c(X)=1})-\gamma\right|^2\right],
\end{aligned}$$

reversing the steps of the proof of Proposition 35. ■

Appendix G. Self-Realization Does Not Imply Equal to Usefulness for All

In the following appendix we provide a short example to show that calibration does not imply that the low swap regret guarantee is equally useful for all. Note that conceptually the statements could seem strange to the reader. The predictor in our case has access to the nature's outcome. This surely is unrealistic. Nevertheless, it does not undermine the argument that a perfectly calibrated predictor exists which is of different usefulness to different downstream decision makers.

For the example we reuse *simple loss functions* (Definition 10. They describe binary-valued elicitable properties on binary outcome sets. The optimal achieveable risk, Bayes risk, for a distribution $P \in \Delta(\mathcal{Y})$ for $\mathcal{Y} = \{0, 1\}$ is given by,

$$\mathcal{B}_q(P) \coloneqq \mathbb{E}_{Y \sim P}[\ell_q(Y, \Phi_q(P))].$$

Lemma 37 (Calibration Does not Guarantee Cost Parity for Arbitrary Downstream Decision Makers with Equal Bayes Risk)**.** *Let $\mathcal{X}$ be finite (with size T), $\mathcal{Y} = \{0, 1\}$. Let ℓ_c and ℓ_d be simple losses with $c, d \in (0, 1)$ and $d < c$. For every choice of c, d there exists a data distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ regular wrt. $\mathcal{P}$ with the $\mathcal{X}$-marginal being the uniform distribution and $Q \in \Delta(\mathcal{Y})$ being a constant conditional distribution $D_{Y|X=x}$ such that*

(i) the Bayes risks are equal $\mathcal{B}_c(Q) = \mathcal{B}_d(Q)$.

(ii) and there exists a calibrated predictor $p\colon \mathcal{X} \to \Delta(\mathcal{Y})$ such that,

$$|\mathbb{E}_{(X,Y)\sim D}[\ell_c(Y, \Phi_c(p(X)))] - \mathbb{E}_{(X,Y)\sim D}[\ell_d(Y, \Phi_d(X))]| \geq C,$$

for some $C \in \mathbb{R}$.

Proof We first define the stationary conditional distribution Q on the binary outcome set via $q \in [0, 1]$. Let $q = \frac{1}{\frac{1-c}{d}+1}$. Observe $d < \frac{1}{\frac{1-c}{d}+1} < c$.

It remains to check the Bayes risk condition,

$$\begin{aligned}
\mathcal{B}_c(Q) &= (1-c)q \\
&= \frac{1-c}{\frac{1-c}{d}+1} \\
&= d\left(1 - \frac{1}{\frac{1-c}{d}+1}\right) \\
&= d(1-q) = \mathcal{B}_d(Q).
\end{aligned}$$

Now, let us define the predictor p.

We define the following prediction function for every $\omega \in \Omega$: if $y(\omega) = 0$, then $p(X(\omega)) = f$ for some $f \in [0, 1]$ such that $d < f < q$. If $y(\omega) = 1$, then with probability

$$x \coloneqq \frac{(1-q)f}{q(1-f)},$$

the prediction is $p(X(\omega)) = f$, otherwise $p(X(\omega)) = 1$. Since $f < q$ and $d, c \in (0, 1)$,

$$0 < x = \frac{(1-q)f}{q(1-f))} < 1.$$

For the second, note that all predictions $p(X(\omega)) = 1$ are calibrated by definition. For the predictions $p(X(\omega)) = f$ we have

$$\begin{aligned}
\mathbb{E}_{Y\sim Q}[Y|p(X) = f] &= \frac{qx}{(1-q)+qx} \\
&= \frac{q\frac{(1-q)f}{g(1-f)}}{(1-q)+q\frac{(1-q)f}{q(1-f)}} \\
&= \frac{\frac{(1-q)f}{1-f}}{\frac{(1-q)(1-f)}{1-f}+\frac{(1-q)f}{1-f}} \\
&= \frac{\frac{(1-q)f}{1-f}}{\frac{(1-q)}{1-f}} \\
&= f.
\end{aligned}$$

So, we can finally show the statement

$$\begin{aligned}
&|\mathbb{E}_{(X,Y)\sim D}[\ell_c(Y, \Phi_c(p(X)))] - \mathbb{E}_{(X,Y)\sim D}[\ell_d(Y, \Phi_d(X))]| \\
&= |\mathbb{E}_{X\sim D_X}[qx(1-c) - (1-q)d]| \\
&= (1-q)\left|\frac{f}{1-f}(1-c) - d\right| \\
&\geq C,
\end{aligned}$$

because $\frac{f}{1-f}(1-c) - d \neq 0$ if $f < q$. ■

In other words, the predictions, even though they are calibrated and hence fulfilling some notions of omniprediction (cf. Section 5.3), do not guarantee that the predictions are equally useful for every decision maker. This holds despite the forecasting task itself being equally difficult as measured by the Bayes risk.

Appendix H. Proofs and Lemmas

Lemma 38. *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $\mathcal{Y}$ is finite and define $A \subseteq \mathcal{X}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor such that $\Gamma(f(x)) = \gamma$ for all $x \in A$. We have,*

$$\Gamma(\mathbb{E}_D[f(X)|X \in A]) = \gamma.$$

Proof Because $|\mathcal{Y}| < \infty$, $f(x) \in \mathbb{R}^{|\mathcal{Y}|}$ for all $x \in A$. Hence, $\mathbb{E}_D[f(X)|X \in A] \in \mathbb{R}^{|\mathcal{Y}|}$. It remains to show that $\mathbb{E}_D[f(X)|X \in A]$ is in the level set $\Gamma^{-1}(\gamma) \subseteq \mathbb{R}^{|\mathcal{Y}|}$. For that, we use the convex indicator function $I\colon \mathcal{P} \to \mathbb{R}$ with $d \mapsto [\![d \in \Gamma^{-1}(\gamma)]\!]$

$$\begin{aligned}
I(\mathbb{E}_D[f(X)|X \in A]) &= [\![\mathbb{E}_D[f(X)|X \in A] \in \Gamma^{-1}(\gamma)]\!] \\
&\leq \mathbb{E}_D\left[[\![f(X) \in \Gamma^{-1}(\gamma)]\!]|X \in A\right] \\
&= 0,
\end{aligned}$$

by Jensen's inequality. Hence, $\mathbb{E}_D[f(X)|X \in A] \in \Gamma^{-1}(\gamma)$. ■

Lemma 39. *Let $\mathcal{Y}$ be finite and Γ be the full distribution property, and thus $\mathcal{R} = \Delta(\mathcal{Y})$. If $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ is a bounded loss function, then ℓ is locally outcome Lipschitz with respect to Γ.*

Proof Let Φ be the property elicited through ℓ. For all $v \in \mathrm{im}\Phi$, $\gamma \in \mathrm{im}\Gamma$, $P \in \Gamma^{-1}(\gamma)$ and $P' \in \mathcal{P}$,

$$\begin{aligned}
&|\ell(P, v) - \ell(P', v)| \\
&= |\sum_{y \in \mathcal{Y}} (P(Y = y) - P'(Y = y))\ell(y, v)| \\
&\leq C|\sum_{y \in \mathcal{Y}} (P(Y = y) - P'(Y = y))| \\
&\leq C \sum_{y \in \mathcal{Y}} |(P(Y = y) - P'(Y = y))| \\
&\leq Cm(\Gamma(P'), \gamma),
\end{aligned}$$

because m is the total variation distance. ■

Lemma 40 (When Decision Calibration is Equivalent to Calibration on Decision Sets)**.** *Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f\colon \mathcal{X} \to [0, 1]$ is a mean-predictor, i.e., it predicts the probability of $Y = 1$. Furthermore, suppose f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$. Let ℓ_q be a simple loss function with parameter $q \in [0, 1]$. In addition, suppose $P_D(f(X) \leq q) > 0$. Then, f is calibrated on the decision sets of ℓ_q, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) \leq q] = 0$ and $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) > q] = 0$[26], if and only if, f is decision calibrated with respect to the loss function ℓ_q.*

Proof Let us rewrite the decision calibration term,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[\mathbb{E}_{Y\sim D_{Y|X=x}}[\ell_q(Y, f(x))] - \mathbb{E}_{Y\sim f(X)}[\ell_q(Y, f(x))]] \\
&= \mathbb{E}_{X\sim D_X}[P(Y = 0|X = X)q[\![f(X) > q]\!] + P(Y = 1|X = X)(1 - q)[\![f(X) \leq q]\!] \\
&\quad - (1 - f(X))q[\![f(X) > q]\!] - f(X)(1 - q)[\![f(X) \leq q]\!]] \\
&= \mathbb{E}_{X\sim D_X}[(P(Y = 1|X = x) - f(X))q[\![f(X) > q]\!] \\
&\quad + (f(X) - P(Y = 1|X = x))(1 - q)[\![f(X) \leq q]\!]] \\
&= q\mathbb{E}_{X\sim D_X}[(P(Y = 1|X = x) - f(X))[\![f(X) > q]\!]] \\
&\quad + (1 - q)\mathbb{E}_{X\sim D_X}[(f(X) - P(Y = 1|X = x))[\![f(X) \leq q]\!]] \\
&=: L(q).
\end{aligned}$$

26. If f matches the average, the first assumption implies the second.

Note that the following two equations hold,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))[\![f(X)>q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&=\mathbb{E}_{(X,Y)\sim D}[(Y-f(X))|f(X)>q]P_D(f(X)>q),
\end{aligned}$$

and,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)\leq q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))|f(X)\leq q]P_D(f(X)\leq q) \\
&=\mathbb{E}_{(X,Y)\sim D}[(f(X)-Y)|f(X)\leq q]P_D(f(X)\leq q).
\end{aligned}$$

With these reformulations at hand, it is easy to see that calibration on decision sets of ℓ_q implies decision calibration with respect to the loss function ℓ_q, because,

$$\begin{aligned}
L(q) &= q\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&\quad+(1-q)\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))|f(X)\leq q]P_D(f(X)\leq q) \\
&= q0P_D(f(X)>q)+(1-q)0P_D(f(X)\leq q)=0.
\end{aligned}$$

The reverse implication requires a bit more reasoning. We argue via contraposition. Let us assume that $\mathbb{E}_{(X,Y)\sim D}[Y-f(X)|f(X)\leq q]\neq 0$. Then,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}[Y-f(X)] \\
&=\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&\quad+\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q) \\
&=0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&=-\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q),
\end{aligned}$$

respectively,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)>q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)\leq q]\!]].
\end{aligned}$$

Now,

$$\begin{aligned}
L(q) &= \mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q) \\
&\neq 0,
\end{aligned}$$

because $P_D(f(X)\leq q)>0$ by assumption. ■

Lemma 41 (Matching Averages by Non-Extremal Predictions I)**.** *Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f\colon \mathcal{X} \to [0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y = 1$. Furthermore, we assume there exists $f_{\inf} \coloneqq \inf \operatorname{im} f > 0$. Then, f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$, if and only if f is decision calibrated with respect to the loss function $\ell_{\frac{f_{\inf}}{2}}$.*

Proof Observe,

$$\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = \mathbb{E}_{(X,Y)\sim D}\left[(P(Y = 1|X = x) - f(X))[\![f(X) > \frac{f_{\inf}}{2}]\!]\right].$$

and

$$\begin{aligned} L\left(\frac{f_{\inf}}{2}\right) &= \frac{f_{\inf}}{2}\mathbb{E}_{X\sim D_X}\left[(P(Y = 1|X = x) - f(X))[\![f(X) > \frac{f_{\inf}}{2}]\!]\right] \\ &\quad + \left(1 - \frac{f_{\inf}}{2}\right)\mathbb{E}_{X\sim D_X}\left[(f(X) - P(Y = 1|X = x))[\![f(X) > \frac{f_{\inf}}{2}]\!]\right] \\ &= \frac{f_{\inf}}{2}\mathbb{E}_{(X,Y)\sim D}[Y - f(X)]. \end{aligned}$$

Since $\frac{f_{\inf}}{2} > 0$, $L(f_{\inf}) = 0$ if and only if $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$. ■

Lemma 42 (Matching Averages by Non-Extremal Predictions II)**.** *Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f\colon \mathcal{X} \to [0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y = 1$. Furthermore, assume $f_{\inf} \coloneqq \inf \operatorname{im} f = 0$, and $|\operatorname{im} f| < \infty$, hence there exists $v \in [0,1]$ such that $v > 0$ but $v < v'$ for all $v' \in \operatorname{im} f \setminus \{0\}$. Then, f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$, if and only if f is decision calibrated with respect to the loss functions ℓ_v and ℓ_0.*

Proof It holds $L(0) = 0$ and $L(v) = 0$. In detail,

$$L(0) = \mathbb{E}_{X\sim D_X}\left[(f(X) - P(Y = 1|X = x))[\![f(X) = 0]\!]\right] = 0.$$

$$\begin{aligned} L(v) &= v\mathbb{E}_{X\sim D_X}\left[(P(Y = 1|X = x) - f(X))[\![f(X) > v]\!]\right] \\ &\quad + (1 - v)\,\mathbb{E}_{X\sim D_X}\left[(f(X) - P(Y = 1|X = x))[\![f(X) = 0]\!]\right] \\ &= v\mathbb{E}_{X\sim D_X}\left[(P(Y = 1|X = x) - f(X))[\![f(X) > v]\!]\right] = 0. \end{aligned}$$

Hence, since $v > 0$, $\mathbb{E}_{X\sim D_X}\left[(P(Y = 1|X = x) - f(X))\mathbf{1}\left[f(X) > v\right]\right] = 0$. Consequently,

$$\begin{aligned} &\mathbb{E}_{X\sim D_X}\left[(P(Y = 1|X = x) - f(X))[\![f(X) > v]\!]\right] \\ &\quad + \mathbb{E}_{X\sim D_X}\left[(f(X) - P(Y = 1|X = x))[\![f(X) = 0]\!]\right] \\ &= \mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0. \end{aligned}$$

■

References

Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, pages 2704–2729. PMLR, 2022.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. In *Uncertainty in economics*, pages 345–375. Elsevier, 1978.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Jarosław Błasiok, Parikshit Gopalan, Linlin Hu, and Venkatesan Guruswami. A Unifying Theory of Distance from Calibration. *Symposium on the Theory of Computing*, 2023.

Federico Cabitza, Andrea Campagner, and Lorenzo Famiglini. Global interpretable calibration index, a new metric to estimate machine learning models' calibration. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 82–99. Springer International Publishing, 2022.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. *Neural Information Processing Systems*, 2024.

Mark H. A. Davis. Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33(3-4):67–93, 2016.

A. Philip Dawid. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

A. Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4): 1251–1274, 1985.

Morris H. DeGroot and Stephen E. Fienberg. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983.

Joseph Diestel and John J. Uhl. *Vector Measures*. American Mathematical Society, 1977.

Paul Embrechts, Tiantian Mao, Qiuqi Wang, and Ruodu Wang. Bayes risk, elicitability, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

Brian S. Everitt. *The Cambridge dictionary of statistics*. Cambridge, 2nd edition, 2002.

Jessica Finocchiaro and Rafael Frongillo. Convex Elicitation of Continuous Properties. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. In *Advances in Neural Information Processing Systems*, volume 34, pages 22046–22057. Curran Associates, Inc., 2021.

Jessie Finocchiaro, Rafael M. Frongillo, and Bo Waggoner. An Embedding Framework for the Design and Analysis of Consistent Polyhedral Surrogates. *Journal of Machine Learning Research*, 25(63):1–60, 2024.

Tobias Fissler and Joanna Ziegel. Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707, 2016.

Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and Nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018.

Rafael Frongillo and Ian A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, December 2021.

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

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792. SIAM, 2024.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *International Conference on Machine Learning*, pages 11459–11492. PMLR, 2023.

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286, 2023.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022a.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 79–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022b.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1983–2026. PMLR, 2024a.

Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. *Advances in Neural Information Processing Systems*, 36, 2024b.

Peter Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 195:47–63, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *International Conference on Machine Learning*, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. *International Conference on Learning Representations*, 2021.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online Multivalid Learning: Means, Moments, and Prediction Intervals. *LIPIcs, Volume 215, ITCS 2022*, 215:82:1–82:24, 2022.

Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. *Advances in Neural Information Processing Systems*, 37:117237–117290, 2024.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Benedikt Höltgen. Practical foundations for probability: Prediction methods and calibration. *philpapers.org*, 2024.

Lunjia Hu and Yifan Wu. Predict to Minimize Swap Regret for All Payoff-Bounded Tasks, April 2024. URL `http://arxiv.org/abs/2404.13503`. arXiv:2404.13503 [cs, stat].

Benedikt Höltgen and Robert C. Williamson. On the richness of calibration. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. ACM, 2023.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment Multicalibration for Uncertainty Estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch Multivalid Conformal Prediction. In *International Conference on Learning Representations (ICLR)*, 2023.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pages 160–171, 2024.

Michael Kirchhof, Mark Collier, Seong Joon Oh, and Enkelejda Kasneci. Pretrained visual uncertainties. *arXiv preprint arXiv:2402.16569*, 2024.

Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-Calibration: Forecasting for an Unknown Agent. *arXiv preprint arXiv:2307.00168*, 2023.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pages 11683–11693. PMLR, 2022.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Nicolas S. Lambert. Elicitation and Evaluation of Statistical Forecasts. *preprint*, 2019.

Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 129–138, New York, NY, USA, July 2008.

Jiuyao Lu, Aaron Roth, and Mirah Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. *arXiv preprint arXiv:2502.12564*, 2025.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: metrics and recalibration. In *Uncertainty in Artificial Intelligence*, pages 1286–1295. PMLR, 2022.

Giacomo Molinari. Trust the evidence: two deference principles for imprecise probabilities. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 356–366. PMLR, 2023.

Allan H. Murphy and Edward S. Epstein. Verification of Probabilistic Predictions: A Brief Review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755, October 1967.

Allan H. Murphy and Robert L. Winkler. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Applied Statistics*, 26(1):41, 1977.

Georgy Noarov and Aaron Roth. The Statistical Scope of Multicalibration. In *Proceedings of the 40th International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.

Georgy Noarov and Aaron Roth. Calibration for decision making: A principled approach to trustworthy ML, 2024. URL `https://www.let-all.com/blog/2024/03/13/calibration-for-decision-making-a-principled-approach-to-trustworthy-ml/`. Accessed 11th of February, 2025.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023.

OpenAI (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774v6*, 2024.

Kent Harold Osband. *Providing Incentives for Better Cost Forecasting (Prediction, Uncertainty Elicitation)*. PhD Thesis, University of California, Berkeley, 1985.

Juan C. Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult Lessons on Social Prediction from Wisconsin Public Schools. *arXiv preprint arXiv:2304.06205*, 2023.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(3):731–817, 2011.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *The Thirty Eighth Annual Conference on Learning Theory*, pages 4937–4972. PMLR, 2025.

Aaron Roth and Mirah Shi. Forecasting for Swap Regret for All Downstream Agents. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pages 466–488, New York, NY, USA, December 2024. Association for Computing Machinery.

Andrew L. Rukhin. Loss functions for loss estimation. *The Annals of Statistics*, 16(3): 1262–1269, 1988.

Kornelius Räth and Nicole Ludwig. Marginal calibration in regression does not guarantee useful uncertainty estimates. *Forthcoming*, 2025.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 52(2): 274–294, 1985.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory*, pages 482–526. PMLR, 2014.

Ben Van Calster, David J. McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 17(1): 230, 2019.

C. Van Fraassen. Belief and the will. *The Journal of Philosophy*, 81(5):235–256, 1984.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR, 2020.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration. *Neural Information Processing Systems*, 2021.