DOI: 10.1017/psa.2025.17

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

Resolving the Reference Class Problem At Scale in Machine Learning

Aaron Roth[1] and Alexander Tolbert[2]

[1]University of Pennsylvania

[2]Emory University

Abstract

We draw a distinction between the traditional reference class problem which describes an obstruction to estimating a single individual probability—which we re-term the individual reference class problem—and what we call the reference class problem at scale, which can result when using tools from statistics and machine learning to systematically make predictions about many individual probabilities simultaneously. We argue that scale actually helps to mitigate the reference class problem, and purely statistical tools can be used to efficiently minimize the reference class problem at scale, even though they cannot be used to solve the individual reference class problem.

[1] Salmon [1977] specifies that reference classes should be "objectively homogeneous" — i.e. chosen such that no further partitioning of the reference class results in a different conditional probability for the event in question, compared to its marginal probability over the whole reference class. But as this is a requirement over all possible partitionings of the data, it is not something that can be found (or even verified) using finite data and computation — a difficulty shared with collective based formalizations of randomness [Martin-Lo¨f, 1966] and probability [Dawid, 1985].

1 Introduction

Statistical inference is fundamental to data-driven decision-making. However, this process has foundational challenges, including the reference class problem (Reichenbach [1949] — see e.g. H´ajek [2007] for a modern treatment). The *reference class problem* refers to the general issue of determining the appropriate reference class over which to estimate probabilities. It comes about from the fact that finite amounts of data do not provide us access to “individual probabilities” (see Dawid [2017]) — we can only estimate aggregates over sufficiently large groupings of the data, or *reference classes* — and the decision of which reference class to aggregate over can dramatically change what our estimates are. Yet, there is not in general any way to uniquely choose a reference class using only finite amounts of data[1].

We make a new distinction between different manifestations of the reference class problem. We call the type of reference class problem that has been traditionally studied the *individual reference class problem*. This is the reference class problem as it arises in the context of making a single probabilistic forecast about an individual: for example, in the context of life insurance, whether a particular individual will die within the next 12 months, or in the context of criminal justice, whether a particular inmate will go on to commit another crime if released on parole.

As these examples might already bring to mind, however, forecasts of this sort are rarely made exactly once: rather they are often systematically and repeatedly made across many individuals. Within the domains of insurance, criminal justice, medicine, technology, etc., increasingly sophisticated statistical and machine learning methods are used to make probabilistic predictions about individuals at scale. A natural concern is that this will correspondingly lead to the *reference class problem at scale* — that the arbitrariness and indeterminacy that the individual reference class problem injects into individual prediction will grow with the scale with which we are now making inferences, and lead to indeterminacy amongst *models* which will result in wide-spread, arbitrary decision making. This concern has been highlighted in the computer science and “algorithmic fairness” literatures under the name of the *predictive multiplicity* or *model multiplicity* problem — see e.g. Marx et al. [2020] and Black et al. [2022].

Our thesis is that while the Individual Reference Class Problem poses significant challenges in justifying predictions about individuals, scale in fact serves to mitigate rather than propagate the problem. When making predictions about many people, there is a way to resolve the Reference Class Problem at Scale: the area of disagreement of two models can itself be used to constructively falsify and improve at least one of the models. The consequence of this is that we can never find ourselves in a situation in which we have two competing "equally well justified" models that make significantly different predictions on a significant fraction of the population. This is in contrast to the individual prediction problem, in which our inability to uniquely pick a reference class can put us in the position of having two "equally well justified" but mutually incompatible estimates of an individual probability. Thus the practical consequences of the reference class problem diminish at scale, although it remains a problem for individual predictions.

More generally, one can worry about a *statistical evidence paradox* that might arise when two different models, f and f', each making probabilistic predictions about some predicate, are equally consistent with the data across all considered statistical tests, including e.g., measures of accuracy and bias across different data divisions, and yet, the two models produce very different predictions. If this paradox arises, it highlights the limitations of traditional statistical evidence because the evidence is consistent with widely varying predictions. Here we highlight *multicalibration* [HebertJohnson et al., 2018] (see also its generalization to "outcome indistinguishability" [Dwork et al., 2021]), a recent framework emerging from the computer science literature on algorithmic fairness as a way to think about and simultaneously enforce consistency of a model with statistical evidence coming from arbitrary collections of reference classes. Multicalibration is intellectually closely related to Dawid's notion of computable calibration [Dawid, 1985] and the collectives-based formalization of randomness originating with Von Mises [1981] and Martin-L¨of [1966]. Through this connection, it also bears a resemblance to classical ways of thinking about the reference class problem, such as Salmon's "Objectively Homogeneous Reference Classes" [Salmon, 1977]. However, it differs in that it focuses on those reference classes that can be identified with bounded amounts of computation and data. This feature makes it actionable in finite data settings that arise in practice. Although, in general, multicalibration (even with respect to all computationally identifiable reference classes) does not imply uniqueness of predictions and so does not resolve statistical evidence paradoxes, any pair of models which witnesses a statistical evidence paradox witnesses a reference class on which one of the two models must be inconsistent, deriving from the region of disagreement of the two models. Cross calibration, in the style of Roth et al. [2023], corresponds to enforcing multicalibration on reference classes that derive from these disagreement regions, which has the effect of eliminating them and resolving statistical evidence paradoxes.

2 The Individual Reference Class Problem

The Reference Class Problem is a fundamental issue in statistical inference [Venn, 1866, Reichenbach, 1949, H´ajek, 2007]. The problem arises when an event or individual can be classified in multiple ways, each leading to a different probability assessment. H´ajek [2007] gives a variant of the following definition, which we modify here to provide formalism that will be useful to our subsequent discussion.

Definition 2.1. *Fix a universe of elements* $e \in \mathrm{X}$ *(corresponding to e.g. records pertaining to people) and a distribution* D *over these elements. Fix a binary predicate*[2] $F : \mathrm{X} \to \{0,1\}$, *as well as a collection of subsets of the universe of elements* $S_1,...,S_n \subseteq \mathrm{X}$ *called* reference classes. *Suppose our access to the underlying probability distribution is limited to sampling from it, and so we can measure conditional probabilities* $\Pr[F(e) = 1|e \in S_i]$ *for sufficiently large reference classes, but cannot directly access "individual probabilities"* $\Pr[F(e) = 1|e]$ *for* fixed $e \in \mathrm{X}$ *(and we might not even make intellectual commitments that posit that these are coherent*[3]*). Then the individual reference class problem arises for an individual element* $e \in \mathrm{X}$ *if there are multiple incomparable reference classes* $S_i \neq S_j$ *such that neither* $S_i \subset S_j$ *nor* $S_j \subset S_i$, $e \in S_i$, $e \in S_j$, *but* $\Pr[F(e) = 1|e \in S_i] \neq \Pr[F(e) = 1|e \in S_j]$.

When the reference class problem arises, there is no unique way to assign $F(e)$ *a probability by estimating a conditional probability conditional on a reference class* S_i.

[2] If we desire, we may also view F as being a mapping $F : \mathrm{X} \to \Delta\{0,1\}$ from universe elements to *random variables* supported on $\{0,1\}$, allowing that the outcome in question might still be stochastic even conditioning on e. This might be the case if e.g. $F(e)$ represents some future (as yet unrealized) outcome for an individual e.

[3] Since we allow that $F(e)$ may be a random variable, we allow (but do not require) that even fixing e, $\Pr[F(e)|e]$ may take values strictly between 0 and 1, in a way that could be consistent with e.g. probabilistic evolution of the universe even fixing all possible current observations. For example, we allow that it might be coherent to speak of a 40% chance of rain tomorrow even when conditioning on all possible current observations.

The reference class problem arises when assigning an “individual probability” to an event when the specific event has only been observed to happen once — or perhaps can only happen once, even in principle. Examples of such individual probabilities include the probability that it will rain tomorrow or the probability that a particular individual will die within the next 12 months. Probabilities for such events cannot be unambiguously estimated from data: in practice we assign the event to an appropriate grouping of “similar” instances in the data, estimate the prevalence of the “similar” events over the grouping, and then impute this estimated probability to the individual in question. For example, in a life insurance scenario, to estimate the probability that a particular individual “Alice” will die within the next 12 months, we might group individuals that we have insured in the past who share demographic similarities — e.g. are of similar ages, genders, weights, have similar medical conditions, etc., and estimate the proportion of people in this grouping who have died within a 1 year period. We will then interpret this proportion as the “probability” that Alice will die within the next 12 months. Training statistical or machine learning models ultimately amounts to finding such groupings automatically, and does not avoid the problem[4]. There is a tradeoff in how we form this grouping or “reference class.” If we insist that the people in the reference class be identical to Alice in every way, we will find that it is empty. So, we must include people who deviate from Alice in a variety of ways in order to have enough samples of data to solve the statistical estimation problem of estimating the 1-year mortality rate within this reference class. But this gives us many degrees of freedom. There are different ways to group people that are “like” Alice in various different ways, and different groupings of people into different reference classes will result in different estimated mortality rates. We can try and justify various choices of reference classes with mechanistic theory of the world, but ultimately any such model must be validated on data, and we may not have unambiguous ways to choose between models (which implicitly assign individuals to reference classes) that are equally predictive.

[4] Buchholz [2023] has argued that in certain cases, the ability of neural networks to learn in high dimensional spaces without overfitting suggests otherwise. We disagree with this assertion—in fact the empirical finding that different, equally accurate neural networks can produce frequently disagreeing predictions [Marx et al., 2020, Black et al., 2022] is strong evidence against it.

The problem of the reference class arises when there are various overlapping categories that an individual may potentially belong to. To illustrate this concern, we examine two cases: the “Gate crashers” case and the Cosmos Club case.

2.1 The Gatecrasher Case

The Gatecrashers scenario was introduced by Cohen [1977] and involves a rodeo event in which some of the attendees failed to pay for admissions — but for any particular attendee, we only have statistical evidence.

We quote here a variant due to Bolinger [2020]. In the variant, among the 1000 people who have attended, only 10 paid admission. We also know that 10 of the attendees are boy scouts, and among this group, 9 of them paid. And 3 of the attendees are Canadians, and among this group, 2 of them paid. Now we are tasked with evaluating the probability that Alfred paid:

> “Alfred is a Canadian former boy scout in F [an attendee at the rodeo]. Let Ga be the proposition ‘Alfred failed to pay’. Conditional on being in F, P(Ga) = .99; conditional on Alfred’s being a former boy scout, P(Ga) = .1; conditional on Alfred’s being Canadian, P(Ga) = .33. With this as evidence, what credence should a rational agent have in Ga? Presumably in asking this question, we think that credences should be substantially constrained by one’s evidence; the problem is that the evidence is not univocal. It supports multiple, competing probability assignments, depending on which reference class we attend to. To identify what credence is rational, we’ll first need to determine whether being a member of F, or former Boy scout, or Canadian, is most relevant to determining whether Alfred failed to pay” [Bolinger, 2020, p. 2420].

The difficulty here is that we do not have any data for the most specific reference class[5] that we could place Alfred into—Canadian former boy scouts (as Alfred is the only one). We can estimate averages over Canadians, or over boy scouts, but those estimates are different, and on what basis are we to say that one reference class is more salient to the task at hand than another? This example demonstrates how the choice of reference class can lead to different probability

[5] The idea that we should choose the “most specific” reference class about which we have reliable data goes back to Reichenbach [1949].

assessments, which could significantly affect the conclusion about Alfred’s payment status. We note in passing that Bollinger seems to be making the assumption that credences should correspond to (frequentist) probabilities here. This is a view that we are sympathetic with, but not one that is universally agreed upon.

2.2 The Cosmos Club Case

The reference class problem is also a common source of stereotyping and racial bias. Racial stereotyping can arise when people make inferences by inappropriately using race to define a reference class. Historian John Hope Franklin recounts an incident at his Washington, D.C., social club, The Cosmos Club, which illustrates this.

> “It was during our stroll through the club that a white woman called me out, presented me with her coat check, and ordered me to bring her coat. I patiently told her that if she would present her coat to a uniformed attendant, ‘and all of the club attendants were in uniform,’ perhaps she could get her coat” [Franklin, 2005, pp. 4, 340].

At the Cosmos Club, the majority of the attendants are black, and there are few black members. This demographic distribution likely led to the woman’s mistaken belief that Franklin, who was a black member of the club, was an attendant.

This case is an example of the reference class problem because the woman incorrectly used the race of the club’s attendants as a reference class to make a prediction about Franklin’s role at the club. She estimated the probability of a person being a staff member, given their race, to be high. If race was the only distinguishing feature, then this might have been a statistically reasonable inference. But in this case, there were other distinguishing features. A more appropriate reference class would have been whether a person is wearing a uniform or dinner attire. This would have led to a more accurate inference [Bolinger, 2020, Gardiner, 2018].

In these cases, the reference class problem comes about because the same event can be classified differently based on the available information. This leads to different probabilities. This issue has significant implications in legal cases, where the determination of guilt or liability often depends on the interpretation of probabilistic evidence [Rhee, 2007]. These are all instances of what we call the *individual reference class problem*, in that the object of interest is always a property or outcome of a single, distinguished individual.

In the next section, we will define the reference class problem at scale, in which the object of interest is a *mapping* from evidence to predictions that can be applied to many individuals. Such an object is the outcome, e.g., of fitting a statistical model.

3 The Reference Class Problem at Scale

3.1 Defining the Reference Class Problem at Scale

The *reference class problem at scale* arises when we systematically make many predictions of individual probabilities. To explore this, we need to study *prediction at scale*. In an informal sense, prediction at scale is conducted using statistical or machine learning models, which are capable of automatically generating predictions about any individual case given descriptive features about that case. These technologies, therefore, implicitly commit to many predictions about individuals, each potentially subject to the individual reference class problem.

Consider a healthcare scenario in which a hospital uses machine learning models to predict patient outcomes—say the likelihood of hospital readmission—for thousands of patients. Such a model would take as input patient records and output for each patient a number between 0 and 1, purporting to be the "probability" that the patient will be readmitted to the hospital within (say) 12 weeks of discharge. Each prediction is of an 'individual probability,' and so depending on the reference class chosen, multiple predictions might be justified as reasonable for each person. The model, however, chooses only one prediction per person. So, one might worry that this indeterminacy of multiple reasonable individual predictions would accumulate into an indeterminacy amongst multiple reasonable *models*, each consistent with the evidence but making very different predictions across a large number of people. Can we, for example, have two equally accurate models, neither of which is statistically falsified[6] by any hypothesis test that we can devise, that nevertheless make very different predictions about a large number of people?

[6] We here speak of "statistical falsification" in the sense of hypothesis testing, not in the sense of logical falsification. Informally, a model is statistically falsified if we can reject the null hypothesis that it is producing forecasts of "true individual probabilities" — or that it is simultaneously consistent with every possible reference class in the sense of Definition 3.1 at some level of confidence.

The possibility of encountering the reference class problem at scale also poses an ethical consideration: if there are multiple very different models equally consistent with the data, on what basis can we justify choosing any one of them to make decisions with important implications about people? To continue our hospital readmission example, once the healthcare system is in possession of the predictions of its model, it might use them to distribute scarce and valuable resources. For example, it might assign patients with the highest predicted probability of readmission-free home visitations by nursing staff. But if two equally well-justified predictive models make very different predictions, then we would also have two methods for allocating scarce and valuable resources to a vulnerable population that result in very different outcomes for many individuals — on what basis can we justify choosing one of the methods over another? To summarize, the individual reference class problem makes predictions about individuals necessarily arbitrary in a certain sense (at least in the practical case in which we must learn from only finite data); is there a way to escape this arbitrariness when making predictions at scale?

Given a universe of elements X, a *model* $f : \mathrm{X} \rightarrow [0,1]$ assigns values $f(e)$ to elements $e \in \mathrm{X}$ that purport to be individual probabilities for some predicate F: i.e., the model purports that $f(e) = \Pr[F(e)|e]$ ("The probability that Alice dies within the next 12 months"). Of course, we should be skeptical of such claims and seek to validate or statistically falsify them based on data. Given samples from the distribution and a reference class $S \subset \mathrm{X}$ that has non-trivial mass under the distribution, we can estimate conditional expectations $\mathrm{E}[F(e)|e \in S]$. Therefore, we can falsify a purported model of individual probability if it significantly deviates from *consistency* with respect to a reference class. Informally, a model is ϵ-consistent with respect to a reference class f if, when we average the model's *predictions* over e in the reference class, we get the same value (up to error ϵ) as we do when we average the actual observed outcomes over the reference class. True individual probabilities $f(e) = \mathrm{E}[F(e)|e]$ would satisfy this property. So a failure to be consistent with respect to a reference class S exhibits an inconsistency of the model with respect to the statistical evidence before us.

Definition 3.1. *A model of individual probabilities* $f : \mathrm{X} \rightarrow [0,1]$ *is* ϵ*-consistent with a reference class* $S \subseteq \mathrm{X}$ *on a predicate* F *if:*

$$\mathrm{E}[f(e)|e \in S] \approx_{\epsilon} \mathrm{E}[F(e)|e \in S]$$

where here $a \approx_\epsilon b$ if $|a-b| \leq \epsilon$. A reasonable "degree of consistency" ϵ will depend on the quantity of data available to us and the frequency of the reference class, which in turn control the accuracy to which we can estimate distributional parameters like $\mathrm{E}[F(e)|e \in S]$.

Just as the individual reference class problem arises when we have an individual to whom there are multiple, seemingly equally good ways to assign probability forecasts as a function of different reference classes, the reference class problem at scale will arise when we have multiple, seemingly equally good models for making predictions at scale (i.e., that are both consistent with the same set of reference classes), that frequently disagree with one another. Informally, we say that the reference class problem at scale arises with respect to a collection of reference classes if *both* models are consistent with all of the reference classes in the set and yet frequently make predictions that substantially differ from one another. More precisely:

Definition 3.2. *Fix a universe of elements $e \in \mathrm{X}$ (corresponding to, e.g., records pertaining to people) and distribution over these elements. Fix a boolean predicate*[7] $F : \mathrm{X} \rightarrow \{0,1\}$*, as well as a collection of subsets of the universe of elements $S_1,...,S_n \subseteq \mathrm{X}$ called* reference classes. *The reference class problem at scale arises if we have two different models $f_1,f_2 : \mathrm{X} \rightarrow [0,1]$ that are both ϵ-consistent on all of the reference classes, but that frequently make substantially different predictions:*

$$\Pr[f_1(e) \not\approx_\epsilon f_2(e)] \geq \epsilon$$

Here $a \not\approx_\epsilon b$ if $|a - b| > \epsilon$. Once again, the parameter ϵ, which we use to measure both consistency with the reference classes and disagreement between the models, should be small, and what a reasonable value is depends on the amount of data we have to estimate the statistical parameters appearing in the definition.

The reference class problem at scale, which we also call a *statistical evidence paradox,* would be paradoxical because it suggests that two models, equally supported by the data, can nonetheless disagree substantially on their predictions. Although not in the language of reference classes, the general problem of having multiple models "equally supported by the data" that

[7] Once again, if we like, we can take $F : X \rightarrow \Delta\{0,1\}$ to map universe elements to *random variables* supported on $\{0,1\}$, allowing the outcome $F(e)$ to be random even conditional on e.

make very different predictions has been noted in the machine learning literature as the *predictive multiplicity* or *model multiplicity* problem [Marx et al., 2020, Black et al., 2022].

An initial question is whether we can reasonably expect to get far enough to encounter a statistical evidence paradox. After all, we have defined one as occurring if we have two models *which are both consistent with all of the reference classes we have considered*, and yet have substantial disagreements about their predictions. Perhaps, just as with the individual reference class problem, it is difficult or impossible to find a model that is simultaneously consistent with many incomparable reference classes. Fortunately, finding a model that is consistent with many reference classes *is* possible — even from finite data. This is known as *multicalibration*, introduced by Hebert-Johnson et al. [2018] and with intellectual roots dating back to Martin-L¨of [1966], Von Mises [1981], and Dawid [1985]. Martin-L¨of [1966] and Von Mises [1981] give a theory of randomness based on "collectives" and Dawid [1985] gives a calibration-based foundation for empirical probability, all of which are based on the idea of consistency with respect to collections of selection rules which can subselect a data sequence based on its observable properties. Selection rules can be thought of as defining reference classes: indeed, Salmon [1977] provided philosophical foundations for proper reference classes, specifying that they should be "homogeneous" — i.e. that it should not be possible to apply a further sub-selection within a reference class in such a way that the conditional probability of the outcome changes. While Martin-L¨of [1966], Salmon [1977], Von Mises [1981], Dawid [1985] were all concerned with the set of all selection rules or reference classes (or countably infinite sets of reference classes, or all computable reference classes) which in the end turns out to give little guidance in settings in which we have only finite data at our disposal, multicalibration as defined by Hebert-Johnson et al. [2018] focuses on collections of reference classes which we can perform statistical estimation over using finite amounts of data[8]

Here, we give an equivalent variant of the definition of multicalibration using the language of reference classes. Note that the reference classes may be arbitrary and may even be defined in reference to the model f (e.g., "the set of all people x such that $f(x) = 0.2$").

[8] Indeed, not just finite amounts of data, but amounts of data and computation that we can control using modestly growing functions of the parameter ϵ governing calibration error.

Definition 3.3. *A model of individual probabilities* $f : \mathcal{X} \rightarrow [0,1]$ *is* ϵ*-multicalibrated with respect to a collection of reference classes* $\mathcal{S}$ *if it is simultaneously* ϵ*-consistent with each reference class* $S \in \mathcal{S}$.

Hebert-Johnson et al. [2018] (see also Roth [2022] for a textbook exposition) show that for any set of reference classes $\mathcal{S}$, it is always possible to find a model f that is ϵ-multicalibrated, with both data and computational requirements that scale reasonably with (i.e. are low degree polynomial functions of) the inverse error tolerance $1/\epsilon$, $\log|\mathcal{S}|$, and $\max_{S\in\mathcal{S}} 1/\Pr[e \in S]$, the inverse of the frequency of the least common reference class. In fact the guarantee is stronger: given *any* model f, it is possible to modify the model (using a modest amount of data) to provide the guarantee that the modified model f' is ϵ-consistent with an arbitrary set of reference classes $\mathcal{S}$, while only improving the squared error of the model.

Definition 3.4. *Fix a universe of elements* $e \in \mathcal{X}$ *(corresponding to, e.g., records pertaining to people) and a distribution* $\mathcal{D}$ *over these elements. Fix a model* $f : \mathcal{X} \rightarrow [0,1]$ *and a binary predicate* $F : \mathcal{X} \rightarrow \{0,1\}$. *The squared error (or* Brier Score*) of the model* f *with respect to* F *is:*

$$B(f) = \mathrm{E}[(f(e) - F(e))^2]$$

As an intermediate lemma, Hebert-Johnson et al. [2018] shows that given any model f, if we discover a reference class S on which f is not ϵ-consistent. It is possible (from small amounts of data) to produce a new model f' that has a lower Brier score (see also Roth [2022] for a formulation closer to what we state here):

Lemma 3.1 (Hebert-Johnson et al. [2018])**.** *Given a model* f *and a reference class* S *such that:*

1. f *is not* ϵ*-consistent on* S
2. S *has probability mass at least* μ_S: $\Pr[e \in S] \geq \mu_S$

then it is possible to efficiently (with an amount of data sampled i.i.d. from the underlying distribution scaling polynomially with $1/\mu_S$ *and* $1/\epsilon$*) produce a model* f' *such that:*

$B(f') \leq B(f) - \Theta(\epsilon^2 \mu_S)$[9]

Let us pause to consider the implications of this lemma, which are several. First, suppose there is a fixed collection S of reference classes that we wish our model to be ϵ-consistent with respect to. One way to achieve our goal is to repeatedly *check* whether our current model fails to be ϵ-consistent with respect to any $S \in$ S, and if so, update the model using the update from Lemma 3.1. Because each time this occurs, the Brier score decreases, and because the Brier score cannot go below zero, this process is guaranteed to halt (after at most $O\left(\max_{S \in \mathcal{S}} \frac{1}{\epsilon^2 \mu_S}\right)$ many iterations) with a model that is ϵ-consistent on every reference class in S — this is essentially the algorithm given by Hebert-Johnson et al. [2018]. Moreover, this process is only accuracy improving—informally, because the "true individual probabilities" would be consistent with respect to *every* reference class, and would also be the global minimizers of the Brier score (as the Brier score is a proper scoring rule), the updates in Lemma 3.1 "march towards truth[10]". Finally, the only way this procedure needs to interact with the data is by estimating conditional expectations over events (reference classes) that have non-trivially large probability, which is a purely statistical problem that can be solved with modest amounts of data. So, it is indeed possible to satisfy the preconditions of a statistical evidence paradox — to find models that are consistent with any collection S of reference classes that we may care to select. But Lemma 3.1 can be applied iteratively *even if we do not commit ahead of time to the reference classes that we will ask for consistency over*, which is what allows us to avoid the reference class problem at scale. In the next section, we describe the "cross-calibration" approach taken by Roth et al. [2023]. Informally speaking, "calibration" asks that a single model be consistent with reference classes defined by its own predictions. Given two models, cross-calibration asks that *both* models be consistent with reference classes defined with respect to both itself and the other model. The cross-calibration approach we discuss in the next section, informally speaking, takes as input two models and, for each, constructs reference classes defined jointly by the predictions of both

[9] Here in writing $\Theta(\epsilon^2 \mu_S)$ we are using asymptotic notation common in mathematical statistics and computer science. In this usage it is merely simplifying the expression by hiding constants — specifying that there exist positive constants c_1, c_2 such that for sufficiently small values of both ϵ and μ_S, we have that $c_1 \epsilon^2 \mu_S \leq B(f) - B(f') \leq c_2 \epsilon^2 \mu_S$ Writing $O(\cdot)$ rather than $\Theta(\cdot)$ indicates the upper bound without the lower bound.

[10] An expression we first heard from Cynthia Dwork.

models. It then asks for multicalibration with respect to these reference classes. This procedure is iterated until a fixed point is reached, and both models are consistent with respect to reference classes defined with respect to both themselves and their counterparts. An important part of the argument is that the fixed point is reached quickly.

3.2 Resolving the Reference Class Problem at Scale

Suppose we were to find ourselves in the presence of a reference class problem at scale: that is, we have two models f_1 and f_2 that substantially disagree substantially frequently, despite both being equally consistent on the collection S of reference classes that we have thought to check. To make this quantitative, suppose that for both $f \in \{f_1, f_2\}$ and every $S \in \mathrm{S}$:

$$\mathrm{E}[f(e)|e \in S] \approx_{\epsilon/2} \mathrm{E}[F(e)|e \in S]$$

And yet

$$\Pr[f_1(e) \not\approx_{\epsilon} f_2(e)] \geq \epsilon$$

The plan will be to construct a new reference class $S(f_1,f_2)$ such that:

1. $S(f_1,f_2)$ is substantially large, and
2. At least one of f_1 or f_2 fail to be $\epsilon/2$-consistent with respect to $S(f_1,f_2)$.

If we can constructively find such a reference class, then not only have we falsified at least one of f_1 and f_2, we can also add $S(f_1,f_2)$ to our set S and perform the update referred to in Lemma 3.1. Since the set $S(f_1,f_2)$ was "substantially large", this update will significantly reduce the Brier score of at least one of the two models, and so (again because Brier scores cannot become negative), this process must converge quickly. But if we are always able to find such a reference class $S(f_1,f_2)$ given any two models that witness a reference class problem at scale with the parameters we have specified, then it must be that when the process halts, the reference class problem at scale has been resolved. Moreover, because every step of this process was accuracy improving, all parties should prefer the models that are produced via this process compared to the models that were input into it: the models that were input into it—unless they survived to be output without modification—were each falsified by some reference class in S (whereas the output models are consistent with all of these reference classes). The models that are output have lower Brier score than all of the previous models produced by this sequence of updates, including the input models — and hence falsify all of the models that precede them.

The problem then reduces to the problem of finding a reference class $S(f_1,f_2)$, given two models f_1,f_2 that satisfy:

$\Pr[f_1(e) \not\approx_\epsilon f_2(e)] \geq \epsilon$

That achieves the two desiderata from above. Here is the construction given in Roth et al. [2023] (a different construction used by Garg et al. [2019] would also work here). Define the "ϵ-Disagreement Region" of the two models $D_\epsilon(f_1,f_2)$ to be the set of points $e \in \mathcal{X}$ such that the two models produce predictions that differ by at least ϵ:

$D_\epsilon(f_1,f_2) = \{e \in \mathcal{X} : f_1(e) \not\approx_\epsilon f_2(e)\}$

By hypothesis, we know that $\Pr[e \in D_\epsilon(f_1,f_2)] \geq \epsilon$. Observe now that we can partition the disagreement regions into two disjoint regions: those points on which f_1 makes a larger prediction than f_2, and those points on which f_2 makes a larger prediction than f_1:

$$D_\epsilon(f_1, f_2) = D_\epsilon^+(f_1, f_2) \cup D_\epsilon^-(f_1, f_2)$$

where

$D_\epsilon^+(f_1, f_2) = \{e \in D_\epsilon(f_1, f_2) : f_1(x) < f_2(x)\}$ and $D_\epsilon^-(f_1, f_2) = \{e \in D_\epsilon(f_1, f_2) : f_1(x) > f_2(x)\}$

We claim that at least one of $D_\epsilon^+(f_1,f_2)$ or $D_\epsilon^-(f_1,f_2)$ satisfy our desiderata.

1. At least one of $D_\epsilon^+(f_1, f_2)$ and $D_\epsilon^-(f_1, f_2)$ must be "substantially large". In particular, as $\Pr[e \in D_\epsilon(f_1,f_2)] \geq \epsilon$ and $D_\epsilon^+(f_1,f_2)$ and $D_\epsilon^-(f_1,f_2)$ form a partition of $D_\epsilon(f_1,f_2)$, we must have that for at least one set $D_\epsilon^\circ(f_1, f_2) \in \{D_\epsilon^+(f_1, f_2), D_\epsilon^-(f_1, f_2)\}$:

$$\Pr[e \in D_\epsilon^\circ(f_1, f_2)] \geq \frac{\epsilon}{2}$$

2. At least one of f_1 and f_2 fail to be $\epsilon/2$-consistent with respect to $D_\epsilon^\circ(f_1, f_2)$. This is because by construction:

$$|\mathbb{E}[f_1(e)|e \in D_\epsilon^\circ(f_1, f_2)] - \mathbb{E}[f_2(e)|e \in D_\epsilon^\circ(f_1, f_2)]| > \epsilon$$

and so whatever value $\mathbb{E}[F(e)|e \in D_\epsilon^\circ(f_1, f_2)]$ takes, we must have for at least one $f \in \{f_1,f_2\}$:

$$|\mathbb{E}[f(e)|e \in D_\epsilon^\circ(f_1, f_2)] - \mathbb{E}[F(e)|e \in D_\epsilon^\circ(f_1, f_2)]| > \epsilon/2$$

Thus, we can choose our reference class to be $S(f_1, f_2) = D_\epsilon^\circ(f_1, f_2)$) and be guaranteed that Lemma 3.1 can be applied so as to decrease the Brier score of at least one of the two models by $O(\epsilon^3)$. Therefore, after at most $O(1/\epsilon^3)$ iterations of this procedure, we have resolved any instance of the reference class problem at scale (up to parameter ϵ). Roth et al. [2023] call this procedure "Model Reconciliation", and the resulting theorem can be formalized as follows:

Theorem 3.1 (Roth et al. [2023])**.** *Fix any $\epsilon, \delta > 0$. There is an efficient algorithmic procedure ("Reconcile") taking as input $O(\ln(1/\delta)/\epsilon^5)$ samples from the distribution that can guarantee the following. Given any two models f_1, f_2, Reconcile outputs a pair of models f_1' and f_2' such that with probability $1 - \delta$:*

1. f_1' and f_2' have only lower Brier score than f_1 and f_2:

$$B(f_1') \leq B(f_1) \quad \textit{and} \quad B(f_2') \leq B(f_2)$$

with the inequalities strict whenever $f_i' /= f_i$ and 2.

f_1' and f_2' almost agree almost everywhere:

$$\Pr[f_1'(e) \not\approx_\epsilon f_2'(e)] \leq \epsilon$$

A brief remark is in order to clarify the parameters of the theorem. A consequence of Lemma 3.1 is that whenever we encounter a reference class that has probability at least ϵ on which a model fails to be ϵ-consistent, we can update the model to reduce its squared error by at least $\Theta(\epsilon^3)$. Because the squared error is bounded between 0 and 1, in the worst case, it starts at 1. By performing these updates, we can drive it down to 0, which would require $O(1/\epsilon^3)$ such iterations (we cannot have more than this without driving squared error to be negative, an impossibility). However, at each iteration, we need to be able to verify from samples that (with confidence $1 - \delta$) at least one of the constructed reference classes $D_\epsilon^+(f_1, f_2)$ and $D_\epsilon^-(f_1, f_2)$ has probability mass at least $\epsilon/2$. This requires $O(\log(1/\delta)/\epsilon^2)$ samples. Multiplying the two bounds gives the bound on the required number of samples in Theorem 3.1. We have chosen to state a simple bound, though it is not the quantitatively tightest bound known. For example, if our initial models f_1 and f_2 are non-trivial, they will not have maximal squared error to begin with: If the squared error of the worst of them $\max(B(f_1),B(f_2)) \leq E$ is bounded by $E < 1$, then the number of iterations improves to $O(E/\epsilon^3)$ and the data requirement bound improves to $O(E \ln(1/\delta)/\epsilon^5)$. If we do not naively take fresh samples at every iteration, but re-use them using so-called “adaptive data analysis” techniques [Dwork et al., 2015, Bassily et al., 2016, Jung et al., 2021] (as was originally done in the analysis of multicalibration by Hebert-Johnson et al. [2018]), then the bound can be further improved to

$O\left(\sqrt{E}\ln(1/\delta)/\epsilon^{3.5}\right)$. We do not wish to focus on the precise dependencies in this bound but to emphasize that it scales only with the error parameters ϵ and δ, and *not* with the complexity of the prediction problem itself. For example, the number of samples needed is independent of how rich the feature space is, so, for example, we can represent individual people with representations *e* consisting of every conceivable piece of information we can record about them without increasing our data requirements. Similarly, our initial models f_1 and f_2 can be arbitrarily sophisticated without increasing our data requirements — in fact, this will improve our data requirements to the extent that it decreases the squared error E of our initial models. To give some sense of the scale of the computation and data requirements, suppose that our initial models have squared error bounded by $E \leq 0.1$, and that we run the algorithm from Theorem 3.1 parameterized to guarantee that with 95% confidence ($\delta = 0.05$), the final pair of models will agree in their predicted individual probabilities up to ±0.05 on 95% of examples ($\epsilon = 0.05$). With these parameters, the number of rounds the algorithm must run before convergence scales as $E/\epsilon^3 = 800$ — something that can be done in seconds on a modern computer — and the number of data points required√

—

scales as $E \ln(1/\delta)/\epsilon^{3.5} \leq 34{,}000$ — a non-trivial but entirely reasonable number of samples in any large-scale prediction problem, and orders of magnitude less than are used to train modern neural network architectures.

Thus, we see that scale actually mitigates the reference class problem: although it does not (and cannot) eliminate the *individual* reference class problem, given modest amounts of data, we cannot have multiple *models* mapping individuals to predictions that are both equally consistent with the data and make substantially different predictions on a substantial number of individuals.

This is because, given two such substantially different models, we have a constructive procedure that is guaranteed to falsify and improve at least one of the two models. Moreover, the quantity ϵ parameterizing the word “substantially” can be driven towards zero by collecting an amount of data scaling polynomially with $1/\epsilon$. Data, therefore, can be used to quantitatively mitigate the reference class problem at scale in a way that it fundamentally cannot be used to mitigate the individual reference class problem. We remark that this theorem does not imply that models trained using standard methods might not disagree substantially on many predictions — indeed, this is known to occur [Marx et al., 2020]. Rather, what Theorem 3.1 implies is that given two such disagreeing models, there is a lightweight procedure (that nevertheless requires a modest amount of additional training data and computing) that can resolve the disagreements in an accuracyimproving way. To return to the healthcare example we used to introduce the reference class problem at scale: if the healthcare system finds itself in possession of two models for predicting hospital re-admission risk that substantially disagree, rather than choosing (arbitrarily) to act on one of them rather than the other, it can apply the model reconciliation procedure to statistically falsify one or both of the models and find more accurate models that rarely disagree. If in the future it finds itself in this position once again, it can iterate the procedure and, therefore, never find itself needing to choose between two equally accurate and well-justified models that nevertheless suggest substantially different downstream actions.

4 Conclusion

In this paper, we have drawn a division between the classical (“individual”) reference class problem, which concerns single predictions, and the reference class problem at scale, which concerns models that systematically map data to predictions. The atomic object of the individual reference class problem is a single prediction; the atomic object of the reference class problem at scale is a model. In both cases, a reference class problem comes about if we have two conflicting atomic objects (substantially different predictions/substantially different models) that are equally consistent with the data before us. Despite the fact that a model is simply a large collection of predictions, and the reference class problem can arise for each of the predictions individually, we have shown that the problem cannot compound across these predictions. So, the reference class problem at scale cannot occur to a substantial quantitative degree when data is prevalent. This means that in scenarios in which we systematically make many predictions (insurance, medicine, etc.), the reference-class problem may have limited bite. We simply cannot find ourselves in a situation in which we have multiple models that substantially disagree on many predictions and are unable to adjudicate between them using purely statistical means.

Acknowledgements

We thank Scott Weinstein, Thomas Grote, and Kasey Genin for their helpful discussions, and Deborah Mayo for first remarking on the similarity between multicalibration and classical work on the reference class problem. A.R. is supported in part by the Hans Sigrist prize, the Simons Collaboration on the Theory of Algorithmic Fairness, and NSF grants FAI-2147212 and CCF2217058.

References

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016. doi: 10.1145/2897518.2897566.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL https://doi.org/10.1145/3531146.3533149.

Ren´ee Jorgensen Bolinger. The rational impermissibility of accepting (some) racial generalizations. *Synthese*, 197(6):2415–2431, 2020. doi: 10.1007/s11229-018-1809-5.

Oliver Buchholz. The deep neural network approach to the reference class problem. *Synthese*, 201 (3):111, 2023. doi: 10.1007/s11229-023-04110-9.

L. Jonathan Cohen. *The Probable and the Provable*. Oxford University Press, 1977. doi: 10.1093/acprof:oso/9780198244127.001.0001. URL https://doi.org/10.1093/acprof:oso/9780198244127.001.0001.

A. P. Dawid. Calibration-Based Empirical Probability. *The Annals of Statistics*, 13(4):1251 – 1274, 1985. doi: 10.1214/aos/1176349736. URL https://doi.org/10.1214/aos/1176349736.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017. doi: 10.1007/ s11229-015-0953-4.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 117–126, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.

2746580. URL https://doi.org/10.1145/2746539.2746580.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 1095–1108, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451064. URL https://doi.org/10. 1145/3406325.3451064.

John Hope Franklin. *Mirror to America: The Autobiography of John Hope Franklin*. Farrar, Straus and Giroux, 2005. ISBN 9780374530471.

G. Gardiner. Evidentialism and moral encroachment. In K. McCain, editor, *Believing in Accordance with the Evidence*, volume 398 of *Synthese Library*, pages 169–195. Springer, Cham, 2018. doi: 10.1007/978-3-319-95993-1 11.

Sumegha Garg, Michael P. Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 809–824, New York, NY, USA, 2019. Association for Computing

Machinery. ISBN 9781450367929. doi: 10.1145/3328526.3329624. URL https://doi.org/10.1145/ 3328526.3329624.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018. doi: 10.48550/ arXiv.1711.08513. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Alan H´ajek. The reference class problem is your problem too. *Synthese*, 156(3):563–585, 2007. doi: 10.1007/s11229-006-9138-5.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees (invited paper). In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 9, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539.
doi: 10.1145/3406325.3465358. URL https://doi.org/10.1145/3406325.3465358.

Per Martin-L¨of. The definition of random sequences. *Information and Control*, 9(6):602–619, 1966. ISSN 0019-9958. doi: 10.1016/S0019-9958(66)80018-9. URL https://www.sciencedirect.com/ science/article/pii/S0019995866800189.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

Hans Reichenbach. *The Theory of Probability*. University of California Press, Berkeley, CA, 2nd edition, 1949. ISBN 978-0520019294.

R. J. Rhee. Probability, policy and the problem of reference class. *The International Journal of Evidence & Proof*, 11(4):286–291, 2007. doi: 10.1350/ijep.2007.11.4.286. URL https://doi.org/10.1350/ijep.2007.11.4.286.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. https://www.cis.upenn.edu/aaroth/uncertainty-notes.pdf, 2022.

Aaron Roth, Alexander Tolbert, and Scott Weinstein. Reconciling individual probability forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 101–110, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593980. URL https://doi.org/10.1145/ 3593013.3593980.

Wesley C. Salmon. Objectively homogeneous reference classes. *Synthese*, 36:399–414, 1977. doi: 10.1007/BF00486104. URL https://doi.org/10.1007/BF00486104.

John Venn. *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to its Logical Bearings and its Application to Moral and Social Science*. Macmillan, London, England, 1866. ISBN 9783337474676.

Richard Von Mises. *Probability, Statistics, and Truth*. Dover Books on Mathematics. Courier Corporation, 1981. ISBN 9780486242149.