The Construction of Algorithmic Predictions in Society

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Benedikt Höltgen
aus München

Tübingen
2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	07.09.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Robert Williamson
2. Berichterstatter:	Assoc. Prof. Dr. Benjamin Jantzen
3. Berichterstatter:	Prof. Dr. Jan-Willem Romeijn

ABSTRACT

As data-driven decision-making is increasingly pervasive in society, algorithmic predictions are often understood as identifying, or at least approximating, properties of the world that are independent of the technical tools; this applies to the technical literature as much as to the public. Individual decisions based on such predictions are then justified by the claim that the models track real relationships between measured attributes. I argue that this view of statistical regularities is based on (i) the common understanding of science as identifying natural laws, (ii) the casino understanding of chances or risks as something that can be identified, and (iii) the fundamental role of statistical regularities in the standard mathematical framework of statistics and machine learning.

In the first part of this thesis, I demonstrate that probabilities are generally constructed and that contrary intuitions rest at least partly on the fact that we cannot directly evaluate them. In particular, I systematise the evaluation of probabilistic predictions through notions of calibration and unify different perspectives on probability by emphasising the ubiquity of model-dependence.

In the second part, I present a constructive perspective on machine learning for societal contexts in particular. I show that the assumption of a data-generating distribution is not necessary for modelling machine learning and that it can be misleading in practice. Theoretically and empirically, I disprove the idea derived from this assumption that fairness considerations necessarily require a deviation from better predictions. I also show that (against received wisdom) the inclusion of demographic attributes can exacerbate inequality and that their current use can be harmful even in auditing for racial discrimination.

In the third and last part, I extend the scope of the foregoing considerations beyond the supervised learning setting. For causal inference tasks, I present a framework dispensing with data-generating distributions, which not only captures established estimators but also paves the way for greater reliability and accountability by making the required assumptions testable. I conclude with a more general critique of common understandings of artificial intelligence.

Taken together, I demonstrate problems with the prevalent understanding of algorithmic predictions and offer conceptual and technical contributions that, in various ways, emphasise their constructed nature and bring theoretical analysis closer to social reality.

ZUSAMMENFASSUNG

Immer mehr Entscheidungen in gesellschaftlichen Kontexten werden automatisiert. Algorithmische Vorhersagen werden dabei häufig so interpretiert, dass sie Zusammenhänge zwischen realen Aspekten der Welt identifizieren (oder zumindest approximieren), die unabhängig von den gewählten technischen Methoden sind. Diese Annahme prägt sowohl die Fachliteratur als auch die öffentliche Debatte und wird oft als Rechtfertigung von Entscheidungen herangezogen, die auf solchen Vorhersagen basieren. Wie ich argumentiere, beruht dieses Verständnis statistischer Regelmäßigkeiten auf (i) einem von Naturgesetzen geprägten Wissenschaftsverständnis, (ii) einem „Casino-Verständnis“ von Wahrscheinlichkeiten als objektiven Größen sowie (iii) der zentralen Rolle, die Gesetzmäßigkeiten in der Statistik zugesprochen wird.

Im ersten Teil dieser Arbeit zeige ich, dass Wahrscheinlichkeiten generell konstruiert sind. Dies wird oft übersehen, da sie nicht direkt evaluierbar sind. Ich systematisiere die Evaluierung probabilistischer Vorhersagen mittels Kalibrierung und vereinheitliche unterschiedliche Auffassungen von Wahrscheinlichkeiten über Gemeinsamkeiten in der Abhängigkeit von Modellen.

Im zweiten Teil entwickle ich ein Verständnis von maschinellem Lernen für gesellschaftliche Kontexte, welches ohne die Annahme von datengenerierenden Wahrscheinlichkeitsverteilungen auskommt, und zeige, dass diese Annahme in der Praxis irreführend ist. Ich widerlege die daraus abgeleitete Vorstellung, dass eine Berücksichtigung von Fairness notwendigerweise mit schlechteren Vorhersagen einhergehen muss, sowohl theoretisch als auch empirisch. Zudem zeige ich, dass—entgegen der gängigen Auffassung—die Einbeziehung demografischer Merkmale Ungleichheiten und Stereotype verstärken kann.

Im dritten Teil erweitere ich den Anwendungsbereich der vorangegangenen Überlegungen. Ich entwickle ein mathematisches Modell für kausale Inferenz, dessen Annahmen und Vorhersagen leichter überprüfbar sind, und schließe mit einer Kritik verbreiteter Auffassungen über künstliche Intelligenz.

Zusammengefasst lege ich Probleme des vorherrschenden Verständnisses algorithmischer Vorhersagen dar und leiste sowohl konzeptionelle als auch technische Beiträge, die den konstruierten Charakter von Vorhersagen aufzeigen und die theoretische Analyse näher an die soziale Realität heranführen.

LIST OF PUBLICATIONS

PUBLISHED

Doh, Miriam, Benedikt Höltgen, Piera Riccio and Nuria Oliver (2025). 'Position: The categorization of race in ML is a flawed premise'. In: *International Conference on Machine Learning*.

Höltgen, Benedikt (2025). 'A thoughtful narrative emerges'. In: *Harvard Data Science Review* 7.4.

Höltgen, Benedikt and Nuria Oliver (2025). 'Reconsidering fairness through unawareness from the perspective of model multiplicity'. In: *EAAMO '25: Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

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

— (2026b). 'The costs of pretending that there are data-generating probability distributions in the social world'. In: *FAccT '26: Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*.

UNPUBLISHED

Höltgen, Benedikt (2026). 'A unifying perspective on probabilities as model predictions'. Preprint.

Höltgen, Benedikt and Robert C Williamson (2026a). 'The accountability gap in causal modelling for automated decision systems'. Preprint.

Tiziano: Allora, capito? È fattibile, fattibile per tutti.
Folco: Cosa è fattibile?
Tiziano: Fare una vita, una vita. Una vera vita, una vita in cui sei tu. Una vita in cui ti riconosci.[1]

— Tiziano Terzani, *La fine è il mio inizio*

1 Roughly: *Did you understand? It is doable, doable for everyone. — What is doable? — Building a life, a life. A real life, a live in which you are yourself. A life in which you recognise yourself.*

ACKNOWLEDGMENTS

I want to first thank Bob for trusting me from the start, for encouraging me to read broadly, and for being a positive influence—especially to stay calm in a hectic research field. A big thank you also to Nuria for inviting me to Alicante, for being an extremely caring and positive supervisor. I would also like to thank all colleagues and especially more senior researchers who have made time to discuss my research ideas and read my drafts (or review this thesis), such as Konstantin Genin, Samira Samadi, Lily Hu, Rediet Abebe, Michael Knaus, Jan-Willem Romeijn, Kate Vredenburgh, and Ben Jantzen.

I am deeply grateful to my parents, who always give a lot and ask for little. I am always aware of how privileged I am to have so much love and support. Equally lucky I am with my siblings, Laura and Phillip, I know we will be there for each other for all of our lives. I am very happy that our family has been growing and now includes Chris and Julius. I am also grateful to Oma Paula for always believing in me, for being an inspiration for perseverance, and for teaching me to read on street signs.

I feel very lucky to still have many friends in Munich who give me the feeling that I still have a home there that will always be waiting for me. I wouldn't know where to begin or end in listing names; you are all important, and my heart fills every time I come back.

Already during my first month in Tübingen (including the trip to Alicante), I found some of my most important pillars here in Chris, Rabanus, and Solveig. Soon after, there were others, especially Sebastian, Nina, Lorenz, and Lara. Sharing an office with friends with similar interests is a real luxury—as is having a friendly and caring research group, including Armando, Charlotte, Laura, Mina, Nan, Rafael, and Wenkai. Other important institutions for me in Tübingen were Lichtenstein, Blauer Salon, and the book club, which met scarcely but still resulted in a number of important friendships. There are many other people who have been important for my time in Tübingen, and I deeply hope (and am confident) that we will keep in touch as we spread from here.

I am also grateful to Nuria's team in Alicante, who made me feel welcome from the start (special shout-out to Aditya and Cristina), and to my flatmates. I want to especially thank Miriam and Piera, who, in between, have become important parts of my life—as close friends even more than as collaborators.

I want to thank all my colleagues, friends, and family who have been (and still are) part of this journey—as fellow travellers, and often also as hosts in Barcelona, Paris, Berlin, New York, Loro Ciuffenna, Ferentino, Alicante, Brussels, Zurich, Savannah, Vancouver, London, Oxford, Lisbon, and, of course, Munich.

CONTENTS

PROLOGUE

A few months ago, in November 2025, the parliament of the German state of Baden-Württemberg (my employer) approved a law that legalised the use of *Palantir*'s data analytics software *Gotham* by the police. For the time being, *Gotham* is only supposed to connect different datasets, as was the case in Norway until the project was discontinued in 2025. In Denmark, however, *Gotham* powers a system for pooling data *and* predicting areas with high risk of crimes (Egbert et al., 2024). *Gotham* has previously also been used for prediction in US cities such as New Orleans, where it was discontinued in 2018 (Stanley, 2018), alongside a ban on predictive policing systems (Stein, 2020).

The rhetoric surrounding machine learning systems is often quite telling of underlying hopes and assumptions, and crime prediction provides particularly suggestive examples. Police in Baden-Württemberg have previously tested a system named *PRECOBS*, referencing the short story (and later Hollywood film) *The Minority Report* where 'pre-cogs' are oracles that can see future crimes. *Palantir*—itself denoting an all-seeing crystal ball—chose the name *Gotham* to conjure a dystopian, crime-ridden city in need of better policing. It is then almost ironic that *PRECOBS* was discontinued in Bavaria because there were too few relevant offences to make the system statistically viable (Wilkens, 2021). Dutch police use the neural network–based *CAS*, which is literally a 'crime anticipation system'.

The very term 'predictive policing' originates from the pioneering company *PredPol*, which has since rebranded as *Geolitica*. According to CEO Brian MacDonald, 'what we do isn't "predictive", what we do is create location-based risk assessments based on historical crime patterns' (Neslen, 2021). Other observers, drawing directly on statistics literature, note that 'predictive policing is a predictive model, which can be described as "a formula for estimating the unknown value of interest"' (Strikwerda, 2021, p. 423). According to popular textbooks, the goal of machine learning 'is to predict t given a new value for x' given 'an input vector x together with a corresponding vector t of target variables', as well as the 'determination of [the joint probability distribution] $p(x, t)$ from a set of training data' (Bishop, 2006, p. 38). But what, then, *are* the relationships between risk assessment, estimation, prediction, and oracles—and how should we think about what these data-driven systems actually do?

Part I

INTRODUCTION

1

CONTEXT

The quiet statisticians have changed our world;
not by discovering new facts or technical developments, but
by changing the ways that we reason, experiment and form our opinions.

— Ian Hacking

1.1 ALGORITHMIC PREDICTIONS

The prediction of crime is one of the most discussed applications of algorithmic predictions. They are used not only for predictive policing technology, but also in the criminal justice system, where individuals' predicted risk of reoffending ('recidivism') can inform decisions about bail or early release. Although such systems also exist in Europe (Andrés-Pueyo, Arbach-Lucioni and Redondo, 2018), the most (in-)famous software is the *COMPAS* system developed and used in the US. The reason for its notoriety is that it has provided the battleground for debates both on comparisons between notions of fairness (Angwin et al., 2016; Chouldechova, 2017; Kleinberg, Mullainathan and Raghavan, 2017) and on the legitimacy of algorithmic predictions in criminal justice (Berk et al., 2021; Wang et al., 2024). While the contributions of this thesis may inform these debates, it will not go into them directly—instead, it will focus on the construction and interpretation of the underlying predictions. In its documentation, the developers state that COMPAS 'is an objective method of estimating the likelihood of reoffending. An individual's level of risk is estimated based on known recidivism rates of offenders with similar characteristics' (Equivant (formerly Northpointe, Inc.), 2017, p. 32).

Another common use case of algorithmic predictions is what is often called the statistical profiling of jobseekers. These systems are widely deployed across Europe (Barnes et al., 2015), often for helping public employment services to allocate job training programmes based on predicted risk of long-term unemployment. While rule-based and logistic regression models are still very common

(Desiere, Langenbucher and Struyven, 2019), some countries are now also developing more complex models based on random forests (Junquera and Kern, 2025) or neural networks (Berman, Fine Licht and Carlsson, 2024). Other applications are the prediction of high-school or college dropout for informing interventions or admissions, respectively (Nielsen et al., 2025; Perdomo et al., 2025).

The tendency to frame policy and allocation problems as prediction problems arguably not only reflects a general fondness for quantification (Porter, 1995) and 'big data' (Boyd and Crawford, 2012), but also an obsession with risk, and individual risk in particular. In the last decades, an increased focus on the anticipation and control of risks has been observed (Douglas, 1992), the result of which has been variously described as a 'risk society' (Beck, 1986) or 'culture of control' (Garland, 2001). At the same time, neo-liberal thinking as well as concerns about stereotyping through too coarse-grained analysis have contributed to an increasing individualisation of risk, in particular in insurance (Krippner, 2023). Algorithmic predictions in society also extend beyond social phenomena like crime, unemployment, or creditworthiness to domains like health. Here, algorithms are used, for example, to predict the risk of diseases or the efficacy of treatments.

1.2 FAIRNESS AND JUSTICE

Algorithmic predictions like in predictive policing have often been taken to be neutral tools even by observers who should know better, as noted by Cathy O'Neil (2014). In her blog post, she argues against the assertion of celebrated journalist Gillian Tett that 'a weather map of crime [is] simply a neutral tool' since 'the algorithms in themselves are neutral' (Tett, 2014). The field of algorithmic fairness, which was founded around that time, was meant to combat this perception (Hardt, 2014). Algorithmic fairness then arguably started its boom with *ProPublica* reporting that the above-mentioned and now infamous COMPAS system was 'biased against blacks' in the sense of unequal error rates of people identifying as black versus white, despite not using this information directly (Angwin et al., 2016). Inspired by this and many similar cases, as well as US anti-discrimination law, a huge literature emerged (and still flourishes) that proposes, criticises, compares, evaluates, and optimises for different notions of group fairness. By now, the field has become fairly mature, with its own big conferences and textbooks.

However, there is also extended criticism against this line of research. In particular, it is now widely accepted that technical fairness notions cannot stand on

their own: As Selbst et al. (2019, p. 59) argue, 'fairness and justice are properties of social and legal systems like employment and criminal justice, not properties of the technical tools within', a view that is echoed e.g. by Birhane et al. (2022b). In a similar vein, Kate Vredenburgh (2022) cautions that 'fairness depends on other standards of justice' and that 'without just institutions, fairness is of little moral worth'. Others go further and argue that algorithmic fairness necessarily only takes the perspective of the decision maker (Kasy and Abebe, 2021) and that it provides a language to 'assimilate outside critiques and concerns into the logics and practices of machine intelligence itself' (Moss, 2021, p. 3). While limited, the fairness perspective can be helpful as a 'diagnostic' (Abebe et al., 2020), to detect discriminatory patterns in algorithmic decision-making; I show in Chapters 3 and 7 that it is necessary and possible to go beyond existing approaches that rely on pre-defined groups such as simplistic race categories that are often assigned from the outside.

1.3 CONSTRUCTION OR DISCOVERY

Sometimes, it is not even necessary to perform fairness audits to find discriminatory patterns. In a recent case, the public employment service in Austria planned to deploy a tool named the *AMS algorithm* for predicting long-term unemployment, in order to inform the allocation of labour market training to job seekers. The plan was to offer job training only to job seekers with a predicted risk of unemployment that was not too high and not too low, as this group was thought to benefit the most. It came under scrutiny especially because of accusations about discrimination, given that women got assigned significantly lower predictions and that care obligations further reduced the predictions, but only for women. In their response to this critique, the agency argued that the algorithmic predictions aimed for 'the *highest objectivity possible* in the sense that the predicted integration likelihood [...] *follows the real chances* on the labour market as far as possible' (as cited by Allhutter et al. (2020), with added emphasis).

The claim to objectivity has been disputed due to the value-laden choices in designing the model, but even critics have granted that it 'merely approximates the labor market's current state' (Allhutter et al., 2020, p. 2). As we have seen above, this rejection of claims to objectivity and neutrality has been central to the algorithmic fairness programme. Still, the quote suggests that there *is* some objective state of the world, some *real chances*, that can be approximated. This is a common perspective in the algorithmic fairness literature. Hardt (2014), as well as later work, consider the issue with neutrality to be rooted mainly in

limited model complexity and limited data. This is a narrative that has long been part of the big data mythology (Boyd and Crawford, 2012) and is still repeated rather uncritically in the algorithmic fairness literature (Roth, Tolbert and Weinstein, 2023). Rather than seeing models as constructed from scratch, even critical work often repeats the phrasing (from a data science textbook, in this case) that statistical models are 'estimating' individual values of interest, such as 'a crime location or offender' (Strikwerda, 2021, p. 423).

As I will argue, this idea (that machine learning models just approximate individual probabilities) contributes to understating specific modelling choices and their effects; as documented by Moss (2021), this is 'how applied machine learning researchers contribute to the idea that numbers can speak for themselves' (p. 21). Consequently, 'the products of machine learning come to instead appear as separate from their labor [...], not only automatic or magic, but also natural and inevitable' (Moss, 2021, p. 13). Indeed, this idea is directly reflected in the predominant theoretical understanding of machine learning.

1.4 MACHINE LEARNING THEORY

The assumption of a 'data-generating process' has become the standard in statistics (see e.g. Figure 1 in the preface of Wasserman (2003)) and 'so much there assumes that the job of statistics is to identify the data generating process' (Vovk and Shafer, 2025, p. 35). This idea is arguably even more deeply ingrained in machine learning, and especially deep learning. A popular textbook by some of the most influential figures in modern machine learning states that '[i]deally, we would like to match the true data-generating distribution p_{data}, but we have no direct access to this distribution' (Goodfellow, Bengio and Courville, 2016, p. 130). Indeed, such distributions are often invoked to understand how machine learning works in the first place. Standard results from learning theory show that samples from a distribution allow us to derive generalisation bounds, that is, guarantee comparable performance on future datapoints. A classic example is Theorem A.2 of Vapnik (1982, p. 170), which he derives using a Symmetrisation Lemma that he calls 'The Basic Lemma' (p. 168):

Theorem 1.1 (Vapnik, 1982)**.**
Take a set of l *datapoints drawn i.i.d. (independently identically distributed) from a distribution* P *over* $\mathcal{X} \times \{0, 1\}$*. Consider a hypothesis class* Λ *of finite VC dimension* h*. Denote by* $U(\alpha)$ *and* $\nu'(\alpha)$ *the expected error rate of* $f_\alpha \in \Lambda$*, and the observed error rate on the training set of* l *points, respectively.*

Then for $k \geqslant l > 2/\epsilon$ *with* $\epsilon > 0$*, we can bound the proportion of draws in which the generalisation error exceeds* ϵ *by*

$$P\left[\sup_{\alpha \in \Lambda} \left|U(\alpha) - \nu'(\alpha)\right| > \epsilon\right] \; < \; 9\frac{(2l)^h}{h!} \cdot \exp\left(-\epsilon^2 l/4\right). \tag{1.1}$$

In practice, beyond the aforementioned temptation of downplaying modelling choices, there are also specific results derived in the standard framework that are questionable at best and misleading at worst. One implication of the standard theory is the fairness-accuracy trade-offs (Corbett-Davies et al., 2017; Hardt, Price and Srebro, 2016; Menon and Williamson, 2018) and their significance. The common idea here is that there is a 'correct' model which largely determines the Bayes-optimal classifier (up to inputs $x \in \mathcal{X}$ with $P(Y = 1|X = x) = 0.5$ in the binary case with symmetric loss), such that any improvement in fairness must represent a loss in accuracy and a deviation from optimality. While these theories could be interpreted as speaking about theoretical concepts (the Bayes-optimal classifier), they are commonly interpreted as applying in some form also to actual models used for decision-making (otherwise, what would be the point of these results?). Indeed, scholars do assert in this context that, for large datasets, such results should hold in practice, that the models 'are likely close to Bayes optimal under the squared loss' (Cruz and Hardt, 2024, p. 2) (echoing Ding et al. (2021)). As the trained models can rarely be considered close to the *empirically* optimal model (otherwise we would not need machine learning), these statements do depend on the existence of the generating distributions and a truly correct model. Recent work, however, shows that it often *is* possible to construct models that are more fair (on some notion) yet equally accurate (Black et al., 2024a; Cooper et al., 2024). In Chapter 6, I demonstrate empirically and theoretically that there is a lot of room for less discriminative models if we account for the distance to the *empirically* optimal classifier, and that there is a complex interplay between data, model class, and choice of attributes. This complexity necessarily gets lost when the analysis is restricted to theoretically optimal models.

There is another way in which the assumption of a true data-generating distribution has had a strong impact on the field of algorithmic fairness. One of the ideas in the field that has taken off more recently is the notion of multicalibration (Hébert-Johnson et al., 2018), which aims to transcend the reliance on specific groupings such as race and gender by optimising for calibration on large numbers of groups. With the common understanding of calibration, this means that for all relevant groups $G \subset \mathcal{X}$ and all predictions $v \in [0, 1]$,

$$P(Y = 1 \mid f(X) = v \wedge X \in G) \approx v. \tag{1.2}$$

This means that the predicted event occurs for 60% of the events in G where the prediction was 60%, and so on. In Chapter 3, I review this concept of calibration and suggest a more general notion, starting from the problem that probabilistic predictions can only be properly evaluated in groups. Now the problem with the idea of multi-calibration is that it derives much of its appeal from the theoretical guarantees derived under the assumption that we can draw ever-more data from the same stable distribution: multi-calibration is thought to bring models ever closer to this true distribution in a way that is largely independent of the chosen groups or the granularity of the input space. As I will elaborate in Chapter 5, the underlying intuition of a 'march towards truth' (attributed to Cynthia Dwork by Roth and Tolbert (2025)) is misguided not only in its stipulation of a true distribution: The idea that the algorithms in question work 'independent of how rich the feature space is', even up to individual people (Roth and Tolbert, 2025) is also misguided, given that they rely on sampling *without* replacement, which does not capture inference beyond the training data if each data point is unique.

While, as machine learning researchers, we have an intuitive idea of what 'i.i.d. sampling from distributions' means, it is difficult to pin down the actual content of it. It is, therefore, instructive to consider its origins. The binary sampling case was already discussed in *Ars Coniectandi* by Bernoulli (1713) and is now called a Bernoulli trial—the intuition for such a sample is the flip of a coin or the drawing from an urn. A crucial step towards drawing from general distributions rather than finite populations (or urns) was then made by Fisher (1922, p. 311) through the introduction of 'hypothetical infinite populations':

> Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. [...] This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample.

The important part is the 'hypothetical' here; for Fisher, 'they do not, of course, correspond directly to reality' (Lenhard, 2006, p. 73).[1] In his own words (Fisher, 1955, p. 71),

> the phrase "repeated sampling from the same population" does not enable us to determine which [hypothetical] population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination.

1 The acceptance of unrealistic modelling assumptions 'as if' they were true is, of course, common in the sciences and not necessarily a problem (Appiah, 2017; Vaihinger, 1911).

Later on, as already reflected in this quote, 'in the standard story [...] the "hypothetical" part of this notion—which is crucial to the concept—gets dropped [...] Fisher did not anticipate the co-option of his framework and was, in large part for this reason, horrified by subsequent developments by Neyman and Pearson' (Kass, 2011, p. 8).

Fisher himself thought that 'the idea of an infinite hypothetical population is [...] implicit in all statements involving mathematical probability' (Fisher, 1925, p. 700). This philosophical idea of hypothetical frequentism has come under much scrutiny in the last 100 years (Hájek, 2009) and is considered not convincing for events that are not repeatable (Dawid, 2017), let alone infinitely often. In Chapter 4, I present a unifying perspective on probability, highlighting the role of models and calibration, without any need for infinities (also clarifying the so-called 'reference class problem' that relates to Fisher's point about the underdetermination of hypothetical populations). Then in Chapter 5, I demonstrate that we do not need the notion of sampling from abstract distributions for machine learning theory (in particular, for results such as Vapnik's), and that, on the contrary, this idea can be harmful.

1.5 FROM PROBABILITY TO CAUSALITY

Another common problem with algorithmic predictions, which is increasingly recognised, is that they do not model the effects of the very decisions they inform (Kern et al., 2025; Liu et al., 2025; Wang et al., 2024). Consider the aforementioned case of unemployment prediction: What these models ultimately aim at is intervening, so decision makers need to know which interventions lead to desirable outcomes. Common approaches target either job-seekers with high risk or, as we have seen in the Austrian case, job seekers with medium risk of long-term unemployment. Rather than predicting the benefits, both of these approaches use heuristics of who may benefit, based on predicted risk, which is often argued to be suboptimal (Zezulka and Genin, 2024).

Indeed, there is a substantial literature in other fields that aims explicitly to model these effects. Especially in statistics, economics, and epidemiology, scholars have developed mathematical frameworks for *causal inference*. The two most popular frameworks today are structural causal models (SCMs) with graphical representations, advocated especially by Pearl (2009) and Spirtes, Glymour and Scheines (2000), and the potential outcomes framework originally proposed by Neyman (1923/1990) and Rubin (1974). The evaluation of labour market programmes in particular has produced a large literature in economics, perhaps most famously the study of LaLonde (1986). For settings of algorithmic decision

making that I am considering, most scholars prefer the Neyman-Rubin models; proponents argue that it allows one to focus more on specific outcomes of interest and accommodate individual problem settings (Imbens, 2020; Markus, 2021). In line with this, Neyman-Rubin models are also suggested by machine learning researchers working on automated decision systems (Liu et al., 2025; Zezulka and Genin, 2024).

The commonalities between the modern form of what I shall call Rubin causal models (RCMs), following (Holland, 1986), and the aforementioned theory of machine learning are striking—perhaps not surprising given their common ancestry in statistics. The starting point is typically the assumption that the observed data are sampled i.i.d. from some distribution—this time a more complex one; for ease of presentation, we will again consider the binary case for both treatment and outcomes. Typically, RCMs assume a joint distribution P over random variables

$$Y_{1i}, Y_{0i}, T_i, X_i \sim P, \tag{1.3}$$

where T_i is the decision variable indicating whether person i is treated or not, Y_{1i} and Y_{0i} denote the outcome for i upon receiving treatment and control, respectively.[2] For the labour market programme example, this means that T_i indicates whether someone gets offered job training, whereas the potential outcome variables may indicate whether they have a job after a fixed time, say, one year. It is assumed that for every person, both Y_{0i} and Y_{1i} are well-defined, i.e., whether they find a job *if* they don't get offered job training and whether they find a job *if* they are assigned job training, respectively. Of course, we can only ever observe the value of one of the two variables for each individual.

In this sense, causal inference faces a similar problem to probabilistic predictions: Individual predictions, in this case individual causal effects $Y_{1i} - Y_{0i}$, cannot be directly compared with their stipulated true values. Yet it also assumes the existence of these quantities, that are defined in the mathematical framework and supposedly approximated, based on which decisions are then meant to be made. In the case of causal inference, the aim is increasingly to make individualised decisions based on 'fully personalized treatment effect estimates' (Athey and Imbens, 2016, p. 7353), by 'using big data for policy problems' (Athey, 2017, p. 483).

2 In his original formulation, Neyman (1923/1990) spoke of finite populations (or farmland plots), but he assumed a 'true yield' in the background of which the observed yield is an instance; in a frequentist manner, he later defined this as the mean of an infinite number of trials under equal conditions (Neyman and Iwaszkiewicz, 1935)—see the discussion by Rubin (1990); thus, through an inversion of the law of large numbers (more on this later), Neyman, in effect, also assumed an underlying distribution, even if that framing only came after Fisher, as discussed above.

In addition to the assumption of i.i.d. samples from a stable distribution, RCMs need to make further assumptions on the distribution, in particular on the assignment process. To perform causal inference beyond randomised trials, the most common assumption is *unconfoundedness* (sometimes also called conditional independence, ignorability, or selection-on-observables):

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp T_i \mid X_i. \tag{1.4}$$

This means that treatment assignment may have been based on X_i but not on any other variables that could give information about the outcome. In other words, the treatment group should be comparable to the control group *per covariate*. The unconfoundedness setting is 'probably the most important one in practice in the modern CI literature' (Imbens, 2020, p. 1163); a potential problem is, however, that 'the unconfoundedness assumption is not directly testable' (Imbens and Xu, 2024, p. 17).[3]

What further confounds the lack of testability in this assumption is that the predictions themselves are not testable, as we observed above. The literature has largely focused on the 'identification' of parameters that define the underlying distribution—'as if a parameter, once well established, can be expected to be invariant across settings' (Deaton and Cartwright, 2018, p. 10). That is, instead of making testable predictions about the future, most work focuses on analysing the past. This arguably impedes progress on a better understanding of the adequacy of common assumptions. While the practice of easy and widespread benchmarking is widely considered important for ML progress (Donoho, 2024; Hardt, 2025), this is much more difficult for causal inference (Vafa, 2024), given its focus on estimation of theoretical constructs in the past. In Chapter 8, I develop an amendment to RCMs which puts predictions about the future into focus and derive empirical versions of the assumptions, to allow more insight into the adequacy of assumptions. I posit that, as in statistics, 'it makes more sense to place in the center of our logical framework the match or mismatch of theoretical assumptions with the real world of data' Kass (2011, p. 1). This is especially so in social and health domains, as our intuitions from gambling or physics may mislead us there into assuming more stability and modularity than these complex systems offer us.

3 Indeed, the question whether it was satisfied in one of the most famous economic studies (LaLonde, 1986) is still under discussion nearly four decades later (Imbens and Xu, 2024).

1.6 SOCIAL PHYSICS

We have briefly touched on the problematic tendency of assuming that estimated parameters or models extend across settings by default. This has arguably to do with our expectation of science as discovering the true laws of nature. This idea can be found e.g. in epidemiology, with textbooks advocating for quick generalisation 'from the actual study experience to the abstract, with no referent in place or time' (Miettinen, 1985, p. 47)—as criticised by Keiding and Louis (2016). With modern statistical machine learning methods, the invocation of abstract data-generating distributions arguably further contributes to this (Chapter 5). Recent examples include e.g. a sepsis prediction tool rolled out across hundreds of hospitals in the US, which was then found to be much less predictive in practice when audited in an independent study; in this case, the company behind the tool stopped advocating a one-size-fits-all version and instead started advising hospitals to make use of their own local data (Ross, 2022). Another example is the use of a generalist recidivism tool used by judges in the US for informing pre-trial detention decisions that was found to be severely miscalibrated in some counties where it was used (Corey, 2019). Lastly, a recent study found that models for predicting the efficacy of schizophrenia treatment which 'performed well in cross validation were little better than chance when predicting outside of the sample in which they were developed' (Chekroud et al., 2024), leading the authors to warn against unwarranted optimism about 'illusory generalizability', which they perceive to be common in such research.

This not only applies to generalisation across populations, but also to assuming stability within the same population. Especially social predictions have the further problem of what Venn (1866, p. 226) called 'suicidal prophecies' and Vovk (2006) called 'Big decision making': The problem of impact of predictions on the decisions, for example of self-fulfilling or self-defeating prophecies, which today is mostly discussed under the term 'performative prediction' (Perdomo et al., 2020). In the predictive policing case, for example, a high estimated risk can lead to more policing in sometimes already over-policed neighbourhoods, leading to more arrests that do not necessarily improve the overall situation (O'Neil, 2017)—or to fewer crimes being committed, as in the short story *The Minority Report* that was later adapted by Steven Spielberg.

More generally, as argued by York and Clark (2007, p. 716), '[social] projections are typically based on an implicit assumption that the social world behaves in a Newtonian fashion and that social conditions are the product of invariant social or natural laws'. More specifically, these projections 'assume that the future is like the past', which can become 'problematic when the highly spec-

ulative nature of this assumption is not properly acknowledged, and the projections based on this assumption become reified' (p. 725). Nancy Cartwright (1999) makes a related point when she says that 'extending our favourite theories to treat new problems of new kinds' is not dangerous only 'when there is good empirical evidence for the promise of the proposed approach' and 'not objectionable so long as we are clear about what we are doing and what it is going to cost us, and we are able to pay the price' (p. 333). She also observes that natural sciences like physics have focused on finding latent quantities that stand in stable relationships to each other, while social sciences typically consider quantities that are of direct interest or easy to measure. And, as she puts it, 'to suppose that there really is some probability measure over [such quantities], you need a lot of good arguments' (p. 325).

The search for laws of nature governing social phenomena has a long history and predates Fisher's introduction of infinite populations by a century. The astronomer Adolphe Quetelet promoted the analysis of social data in similar ways as natural sciences like astronomy after observing a 'frightening regularity with which the same crimes are reproduced' (Porter, 1986).[4] This 'astronomical conception of society' (Hacking, 1975) regarded observations as samples around a mean—invoking an inversion of the law of large numbers more explicitly than Neyman later on. For both physiological and social measurements, he regarded 'deviations from the mean height not as innate diversity, but as error' (Porter, 1986, p. 68). This meant that there had to be an entity from which the errors were measured, which he then famously called the 'average man' (*homme moyen*), and considered 'specific to every individual nation and also to every individual historical epoch' (Mosselmans, 2005, p. 568). He did consider more fine-grained data, however, which led him to proclaim a 'propensity to crime' (*penchant au crime*) in referring to 'the probability that an individual will commit crime (taking the influence of season, climate, gender and age into account)' (Mosselmans, 2005, p. 568). This effectively meant that he thought of society as a constant urn comprising individuals with a constant propensity of crime.[5] While contemporary and later scholars criticised this model as too simplistic, we have arguably come full circle now with our frameworks that assume true distributions over covariates and crimes, from which we claim to sample i.i.d.! Quetelet's objective

4 Quetelet called his endeavour 'social physics', a term he borrowed from the work of August Comte—who then started using the term 'sociology' as he disapproved of Quetelet's work (Hacking, 1990, p. 39).

5 Desrosières (1997) writes that *il y a, dans l'urne de la société, des individus dotés d'une propension au crime, qui, avec une probabilité constante, manifestent ce fâcheux penchant, a modèle de la cause constante macro-déterministe, résultant de la combinaison d'un "grand nombre de causes identiques, petites et indépendantes"*.

and frequentist understanding of probability left a lasting mark on the nascent field of statistics (Desrosières, 1997), which we have seen resurface in the ideas of Fisher and Neyman.

One scholar building on Quetelet's work was Wilhelm Lexis, who investigated the stability of the purported social laws. Using urn models, he set out to analyse whether the observed regularities in crimes or suicides were stronger than one may expect from draws from a larger urn, that is, whether there was indeed something like a social law that held the conjured urn constant over the years (as Quetelet suggested). He came up with what came to be known as the 'Lexis ratio', measuring the rate of dispersion, that is, comparing the observed stability with the stability that would be due to random fluctuations. What he found, then, was that empirical rates for well-defined populations were significantly *less stable* than random fluctuations. In modern terms, this means that even the marginal expected outcome $\mathrm{E}[Y]$ was not stable (let alone the whole distribution). Lexis came to be 'persuaded that, with respect to moral actions, human societies are fundamentally diverse, or heterogeneous, so that even at a specified moment no single value could be assigned as the probability that any given individual will marry or commit a crime' (Porter, 1986, p. 252). Indeed, leading scholars some 40 years later, such as Keynes and Cherlier, considered his work (published in 1877) 'the first essential step forward in mathematical statistics since the days of Laplace [around 1820]' (Rietz, 1924). A refined version of his work (Coolidge, 1921) essentially contained Fisher's groundbreaking work on ANOVA, as demonstrated by Rietz (1932).

The look into the history of quantification in the social world sheds light not only on how it relies on and reinforces overblown optimism about this world's stability, but also on how it can be used to propagate pseudo-scientific racial prejudice. Many of the statistical methods of the time were developed for categorising species and races, often serving eugenicist agendas. Moss (2021, p. 246) illustrates the ubiquity of eugenicist thinking in early statistics with the example that Fisher first published the iris dataset—still used widely in statistical education—in the *Annals of Eugenics*; Pearson and Galton even held a chair for Eugenics at UCL. Galton, a towering figure in both statistics and eugenics, was strongly influenced by the work of his half-cousin Charles Darwin—although the latter already noted the arbitrariness involved in distinguishing varieties and even species. Venn (1866, p. 42), for one, found Darwin's insights incompatible with Quetelet's ideas of stable types of a nation and notes that 'no one who gives the slightest adhesion to the Doctrine of Evolution could regard the type, in the qualified sense of the term, as possessing any real permanence and fixity'. The idea of distinct human races, which originated in the justifica-

tion of slavery, has been comprehensively discredited in science over the course of the 20th century (Smedley and Smedley, 2005); as I discuss in Chapter 7, modern practices in ML, particularly in algorithmic fairness, bear the danger of contributing to its resurrection.

More generally, it may be observed that our ideas of stability are strongly influenced by settings like casinos (in the case of probability) and physics (in the case of natural laws), and quantification strengthens these associations. As put by Porter (1995, p. 11): 'The credibility of numbers [...] is a social and moral problem. This has not yet been adequately appreciated.'

1.7 INDIVIDUAL RISK

Many of the points raised above come together in one question: How should we understand algorithmic predictions? Common parlance often suggests that they approximate people's actual attributes, and indeed come close to them given enough data and model complexity—they allow a 'march towards truth'. On this view, data points are instances of true distributions that describe the real data-generating processes, and individual predictions track people's individual risk. The works presented in this thesis challenge this narrative in various ways. A central point is to convince the reader that there is no such thing as true individual risk in the background.

Fisher was well aware of the theoretical nature of the infinite populations he introduced; in his view, statistical methods could only summarise data, not provide individual risk:

> The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. [...] Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by any actual sample is irrelevant. (Fisher, 1922, 311f)

One problem, which was clear at least since Lexis, is that the social world does not follow stable regularities as the physical world does. This already holds for coarse-grained properties, and even more so for finer-grained ones. In fact, the assumption that new data would simply fall under the same model was one of the fundamental critiques Fisher raised against Jerzy Neyman and Egon Pearson's approach (Fisher, 1955). Another problem is that we can only ever take some of the information into account—we choose to measure certain attributes

of a person or situation. This is not a practical problem that we can improve on but a fundamental aspect of probability itself (Chapter 4).

Now, it is not a new observation that different modelling choices lead to different predictions. Still, the importance of modelling choices is often downplayed such that resulting predictions appear authoritative; 'it is exercising deontic power by obligating users to follow the reasons machine intelligence provides for their actions, rather than their own' (Moss, 2021, p. 238), that is, it provides 'desire-independent reasons for actions' (Searle, 2006, p. 19). I argue that this is directly linked to the idea *approximating* 'real chances' (Allhutter et al., 2020), of *estimating* 'an individual's level of risk' (Jacobs and Wallach, 2021, p. 382). If we instead see predictive models as constructing predictions 'from scratch' (Chapter 5), it also becomes clear that the phenomenon of model multiplicity—models having similar accuracy but different predictions (Chapter 6)—is a fundamental conceptual issue that does not simply disappear with more data (from the same stable distribution) and more complex models.

In this context, I can now outline some connections among the different chapters of this thesis before proceeding to summarise their contributions individually. We can judge predictions by how well the model performs on sets of events (Ch. 3) and by what effects it has on the world when deployed, but it is important to keep in mind that statistical models are designed to perform well on aggregate evaluations (Recht, 2025b). What makes it so difficult to criticise probabilistic (Ch. 5) and causal (Ch. 8) predictions is that there is no ground truth that they can be compared against. Algorithmic predictions should be seen as numbers that can (at best) be useful for *aggregate* decision making (Ch. 4), and not as measurements of real properties in the world, let alone in individual people. So it is crucial to always scrutinise modelling choices and assumptions (Ch. 6), to be wary of the categories we impose on people (Ch. 7), to keep in mind that these models serve a specific purpose, and to ask which decisions (and which stakeholders) benefit from algorithmic predictions (Ch. 9).

2

SUMMARY OF RESEARCH CONTRIBUTIONS

"Begin at the beginning," the king said very gravely,
"and go on till you come to the end: then stop."

— Lewis Carroll, *Alice in Wonderland*

In this chapter, I provide an overview of the research contributions presented in this thesis. Each section summarises one manuscript I have (co-)authored, describing its respective (personal and scientific) motivation and key contributions. **The manuscripts themselves are included below as Chapters 3 to 9.** They are ordered thematically and roughly chronologically (insofar as it is possible to speak of a chronology for partially concurrent projects of varying duration). The contributions of individual authors are reported at the beginning of each chapter.

The LaTeX files[1] of the published and submitted manuscripts have been integrated largely without alteration, with the following exceptions:

- Text and equations have been adapted where a change in line width or citation style made it necessary.
- The (usually alphabetic) labellings of the appendices have been integrated into the chapter numbering.
- The sections in Chapter 9 also received numbers, with the first section now designated as the introduction.
- The notation in Chapter 3 has been slightly adapted to conform with the notation in the other chapters: $p^\star$ and p replace p and $\hat{p}$, respectively.
- The bibliographies have been pooled into a single one, and entries have been updated to the latest version.

1 This thesis was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. `https://bitbucket.org/amiede/classicthesis/`

2.1 CONSTRUCTING EVALUATION METRICS FOR PROBABILITIES

This section motivates and summarises Chapter 3, 'On The Richness of Calibration' (Höltgen and Williamson, 2023).

2.1.1 *Personal Motivation*

Before starting my PhD, I had already come to the conclusion that probabilities are useful insofar as summing them up for a relevant set of events is predictive of how many of those events occur.[2] That is, if for each event $i \in \mathcal{I}$ we denote by p_i its probability and by $y_i \in \{0, 1\}$ whether it occurs, we need

$$\sum_{i \in \mathcal{I}} p_i = \sum_{i \in \mathcal{I}} y_i \quad \text{or, equivalently,} \quad \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p_i = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y_i. \tag{2.1}$$

Dawid (2017) then taught me this criterion can be called 'calibration' on $\mathcal{I}$. Following the impossibility results of Chouldechova (2017) and Kleinberg, Mullainathan and Raghavan (2017), the idea of 'group-wise calibration', in the sense of calibration on sets of equal prediction per demographic group, was still widely discussed in the fair-ML literature. With Bob, we wanted to explore fairness notions that do not depend on a few pre-defined groups.[3] This dependence has also been criticised in the literature on 'multi-calibration' (Hébert-Johnson et al., 2018); here, the idea is to extend group-wise calibration to more groups, for example, to all efficiently computable ones.

It is often thought that 'calibration also means that by default, we should expect our models to faithfully reflect disparities found in the input data' (Barocas, Hardt and Narayanan, 2023, p. 21). In Chapter 6, I demonstrate that this is not the case for classification disparities, and highlight the dependence on modelling choices. Here, Bob and I aimed to clarify how calibration itself comes with many design choices. We thus decided to write a paper that explores the history and the richness of calibration, in order to carve out the implicit choices in common calibration scores and suggest alternatives. Most notably, it is typically assumed without questioning that (multi-)calibration necessarily requires grouping inputs of similar prediction; one researcher explicitly told me that the appeal of this derived from calibrated predictors then being 'coarsenings of the true conditional distributions'. More generally, the common understanding of calibration depends on assumptions about stable relative frequencies

2 I elaborate on this perspective on probability in Chapter 4.

3 Works that particularly motivated this latter part were the works of Hu and Kohler-Hausmann (2020) and Gabriel Tarde as discussed by Latour (2015), highlighting the constructed interdependence of individuals and groups (see also Chapter 7).

(Dawid, 1985) and, thus, true probabilities. I argue against these assumptions in Chapter 4 and elaborate on the choice of the input space in Chapter 5.

2.1.2 *Summary*

In Chapter 3, we develop a foundational perspective on calibration as an evaluative concept that necessarily depends on choices of grouping and aggregation. While calibration has a local ideal in the standard theory, it always needs to be evaluated through groups, making grouping choices fundamental rather than secondary. We identify three largely independent knobs for selecting groupings: The basic *principle* determines a notion of vicinity, typically in the input space or output space; the *size* of the groups trades off informativeness with robustness of the group-wise errors; the *structure* decides, for example, whether the chosen groups overlap or form mutually exclusive bins. Given such a grouping, there are different ways of agglomerating group-wise errors, common choices being the average or the maximum.

We analyse these aspects mathematically, show how they capture existing notions of calibration, and highlight possible problems such as diagnosing stereotyping or miscalibration on minority groups. Based on these considerations, we introduce global calibration-based fairness scores based on signed errors and risk measures. Overall, we analyse the possibilities and problems of evaluating probabilistic predictions through an analysis of the notion of calibration from the ground up.

2.1.3 *Key Contributions*

- We organise the space of possible calibration scores through a small number of largely independent choices, pointing out considerations and trade-offs relevant to these choices. We capture salient aspects of these trade-offs mathematically.
- We propose the (to our knowledge) first provably consistent estimators of individual calibration, based on kernels and k-nearest neighbours.
- We characterise the space of functions suitable for agglomerating calibration errors into global scores of prediction quality, capturing existing notions of calibration.
- We propose and axiomatise agglomeration functions for global calibration-based fairness scores and show that through suitable choices of grouping,

such scores can provide notions of (sub-)group fairness as well as notions of individual fairness.

- We show that calibration admits a substantially broader interpretation than is standard in the literature, and that different calibration-based evaluations correspond to different epistemic and normative commitments rather than interchangeable ways of operationalising 'individual risk'.

2.2 CONSTRUCTING PROBABILITIES

This section motivates and summarises Chapter 4, 'A Unifying Perspective on Probabilities as Model Predictions' (Höltgen, 2026).

2.2.1 *Personal Motivation*

I had been interested in the philosophy of probability since my first master's programme, writing term papers on the general model-dependence of probabilities. As mentioned above, I later realised of the importance of calibration, providing the second key ingredient to my current understanding of probability. In the PhD, I was then compelled to take up the subject again when noticing its relevance for understanding algorithmic decision making, as described in Chapter 1. It further encouraged me that the importance of better understanding probability was acknowledged by the likes of Cynthia Dwork, who said that the meaning of probability was a 'pressing question' (Burhanpurkar et al., 2021), and that 'without an answer to this definitional question, we don't even know what it is that the ideal algorithm should satisfy' (Dwork, 2022).

Indeed, the widespread idea that machine learning predictions aim to approximate data-generating distributions (Chapter 5) is somewhat analogous to the widespread belief in philosophy that there are epistemic and aleatoric probabilities.[4] Also in philosophy, it is thought that the former ideally match the latter, as captured by the 'Principal principle' (Lewis, 1980). The duality of epistemic versus aleatoric probabilities and the debates between frequentism and Bayesianism are commonplace in both philosophy and machine learning/statistics. In my view, however, there is no such fundamental distinction between kinds of probabilities, and both descriptions are flawed as general accounts of probability. Partly because of its relevance for the understanding of machine learning systems in society, and partly out of my earlier interest,

4 In reference to this duality, probability has been described as 'Janus-faced' (Hacking, 1975).

I set out to clarify that all the probabilities we encounter are indeed model-dependent.[5]

2.2.2 *Summary*

In Chapter 4, I present a unifying account of probability based on the insights that all probabilities are model-dependent and that calibration makes probabilities useful. I introduce the notion of prediction methods which consist of an abstraction scheme as well as a way of turning that abstraction into a prediction. Common cases are symmetry- or frequency-based prediction methods, which idealise, e.g., a gambling setting to the symmetric shape of the device or a medical risk assessment of a person to the presence of relevant genetic markers. But why do we construct probabilities in the first place, if not to approximate true probabilities in the world? The answer is basically that we can often predict how many events of a relevant set will occur, and thereby decide which policy leads to a desirable set of outcomes. In particular, if we are calibrated (in the general sense of (2.1)) on sets of events that are equally valuable to us, then we can predict the distribution of value (utility) that we will receive.

However, the problem of induction implies that we can never know for sure whether we will actually be calibrated. But if we accept that there are certain stabilities that allow us to reliably make predictions in practice, this also means that calibration is possible; a special case of this is the stability of the relative frequency of a class of events. The importance of calibration also explains the desirability of Kolmogorov's axioms: As I demonstrate, the latter guarantee calibration for certain sets of events. I show that this understanding of probability also clarifies the intuitive appeal of other interpretations of probability, such as Bayesianism, frequentism, and propensity accounts. The notion of prediction methods also elucidates that the so-called 'reference class problem' is broader than usually thought, including not only the abstraction step but also the way of processing abstractions. My pragmatic account shows, however, that this does not actually need to be seen as a 'problem': It simply describes the freedom to choose among different methods.

5 Despite the supposed clarity of the key ideas from the outset, this paper has undergone substantive changes from the first draft, as many pairs of eyes have passed over it; in my opinion, the draft has improved significantly thanks to the generous feedback.

2.2.3 *Key Contributions*

- I unify what are usually considered distinct kinds of probability by demonstrating their common dependence on prediction methods and their usefulness via calibration.
- I explain that calibration is often possible, based on combinatorial considerations; this is not meant as a solution to the problem of induction but as a clarification of the role of probability in inductive inference.
- I prove that the probability calculus is required to ensure calibration in certain settings, thereby explaining why its axioms are desirable.
- In line with many normative theories, I show that policies and sets of predictions should be put to scrutiny rather than individual decisions or predictions,
- I connect my account with other interpretations of probability by recovering key intuitions behind them.

2.3 CONSTRUCTING MACHINE LEARNING MODELS: THEORY

This section motivates and summarises Chapter 5, 'The Costs of Pretending That There Are Data-Generating Probability Distributions in the Social World' (Höltgen and Williamson, 2026b).

2.3.1 *Personal Motivation*

I stated above that contemporary machine learning discourse was part of the motivation for my work on probability reported in Chapter 4); this relevance is, however, not straightforward. After giving valuable feedback on a draft of it, my friend Jannik Thümmel asked something like, 'this is all nice and interesting, but what should we do with it now?' Around that time, ICML introduced its position paper track, and so I decided to make the case for a more careful treatment of probability in machine learning—something that Bob had already been advocating for a while. As put by Xiao-Li Meng (2022, p. 341),

> we typically conceptualize that the data at hand is a realization of a generative probabilistic mechanism given by nature or God. [...] We do so regardless of whether we believe or not that the world is intrinsically deterministic or stochastic...

While there are points to be made that apply to machine learning and statistics in general, the issue is particularly pressing and clear when we make predictions about people. As discussed in Chapter 1, the rhetoric around machine learning is especially important when used to justify decisions about people, and the implicit stability assumptions are weakly supported. Early versions of this manuscript made the general points, only highlighting the social setting as a particularly pressing case. But some conference reviews, as well as discussions at workshops where I presented it, made it clearer to me that it was better to focus more on the social setting. Building on Chapter 4 as well as Bob's independent criticisms of data-generating distributions, this chapter argues that data-generating distributions can be misleading and obscure both the choices made and the goals pursued in machine learning practice. Coincidentally, Ben Recht (2025a) recently made similar points in his blog, noting that 'as I dug into the roots of the misunderstandings and misinterpretations of machine learning, I kept hitting against the unassailable belief in the data-generating distribution.'

2.3.2 *Summary*

In Chapter 5, we critically examine the widespread assumption in machine learning and statistics that data are generated by a true underlying probability distribution. We focus on prediction tasks in the social world. We argue that, in such settings, the notion of a data-generating distribution is not just an idealisation—as many models in science are— but that it is a bad one: misleading and unnecessary. We first argue, based on Chapter 4, that probabilities are generally constructed rather than discovered, and that they depend on the chosen abstraction and other modelling choices. We demonstrate that in the case of machine learning, one cannot solve this simply by pretending that the invoked probability distribution is the empirical distribution of the entire population: There are still too many choices left open, such as the chosen attributes and time frame. We illustrate this latter point empirically, highlighting the dependence of individual predictions on the considered time frame.

We then turn to machine learning theory and demonstrate that a central motivation for assuming data-generating distributions, modelling generalisation, actually does not require it and can be done with a more parsimonious framework that models the populations directly. In particular, we derive a generalisation bound analogous to Theorem 1.1 in the alternative framework. Lastly, we analyse how the presumption of a generative distribution shapes the interpretation of machine learning models in practice, particularly by suggesting that learning is a process of approximating true conditional probabilities. We

argue that this framing obscures the task-dependent and choice-laden nature of modelling, contributes to overestimating the significance of theoretical fairness–accuracy trade-offs, and downplays the significance of predictive multiplicity. Overall, we conclude that avoiding the assumption of data-generating probability distributions in social settings leads to a more felicitous understanding of what machine learning models do and what their predictions depend on.

2.3.3 *Key Contributions*

The chapter's key contributions are as follows:

- We demonstrate that the assumption of a true data-generating distribution (Gen-D) in social domains lacks ontological and epistemic justification, even as an idealisation.
- We prove that standard learning-theoretic results do not depend on the assumption of Gen-Ds and can, instead, be derived for finite populations.
- We elucidate how the Gen-D assumption obscures the dependence on particular populations, and illustrate this dependence empirically.
- We show that disagreement between equally accurate models is not merely a finite-sample or optimisation artefact, as often claimed, but a fundamental aspect of modelling.
- We contribute to the literature debunking ideas of objectivity by drawing attention to, and refuting, the sometimes explicit argument that machine learning approximates true probabilities.

2.4 CONSTRUCTING MACHINE LEARNING MODELS: PRACTICE

This section motivates and summarises Chapter 6, 'Reconsidering Fairness Through Unawareness from the Perspective of Model Multiplicity' (Höltgen and Oliver, 2025).

2.4.1 *Personal Motivation*

In preparing my research visit to Nuria's lab in Alicante, we discussed possible research directions for my stay there. My idea was to work on empirical aspects of the ideas discussed in Chapter 5, while Nuria suggested working on

algorithmic fairness, connecting more closely to other work in her lab. I had been interested in model multiplicity—the phenomenon that there are typically multiple models with equal accuracy that disagree on individual predictions—for a while and was supervising a master's thesis about it at the time. Nuria and I agreed to look for modelling choices that affect fairness properties without changing average accuracy. This would serve to illustrate the conceptual points in Chapter 5, and help to debunk claims that fairness considerations necessarily lead to a divergence from the closest approximation to truth. It would also challenge justifications of discriminatory patterns based on claims that a model 'follows the real chances on the labour market as far as possible' (as cited by Allhutter et al. (2020) and discussed in Chapter 1). As it turned out, that very summer, Black et al. (2024a) published a paper making the case that such searches for 'less discriminative algorithms' should be pushed more.

I quickly found evidence that including demographic attributes like gender—as in the use case discussed by Allhutter et al. (2020)—can increase inequality between the two considered genders. Meanwhile, this inclusion did not markedly improve accuracy in the data, and sometimes even decreased it. I also found similar effects when considering self-reported race, which is standardly collected in the U.S. Census data I investigated (I discuss caveats of using race as an attribute in Chapter 7).[6] I conducted theoretical analyses to investigate this counterintuitive effect, deriving results that do not assume being close to the optimal predictor—in contrast to much of the existing literature (Cruz and Hardt, 2024; Hardt, Price and Srebro, 2016). To narrow the gap between theory and reality, I also sought to derive results that are specific to a model class.

2.4.2 *Summary*

In Chapter 6, we empirically and theoretically show that using a group attribute like gender or race can make machine learning predictors more unequal with respect to that grouping, without making predictions more accurate. We first demonstrate theoretically that, depending on the threshold, disparate impact can exceed base-rate differences—that means, the classification rates of the model can be more unequal than the label rates in the data. We show why this can happen particularly easily for logistic regression models using group attributes and demonstrate this phenomenon empirically. We derive bounds on the differences in disparate impact within the class of equally accurate models and

6 In a real case, the Dutch software *CAS* mentioned in the Prologue used the number of citizens with non-Western-born parents for predicting crime hotspots; the attribute was later dropped as it was not seen to improve performance (Oosterloo and Schie, 2018).

prove that these bounds are tight, i.e., they can be achieved in some settings. Our results highlight that the bounds strongly depend on how much mass of the empirical distribution is on inputs with label rates close to 50%.

We then demonstrate empirically that removing a group attribute can actually decrease a model's disparate impact significantly (by up to 80%) while hardly reducing accuracy and sometimes even increasing it. We conduct these experiments on some of the largest available social datasets, `ACS Income` and `ACS Employment` (Ding et al., 2021), as well as Austrian labour market data (Lechner et al., 2020). As group attributes, we consider race and sex for the U.S. datasets, and citizenship and gender for the Austrian one; we test both logistic regression models and gradient boosted trees. Lastly, we investigate a real-world use case inspired by the Austrian labour market algorithm discussed by Allhutter et al. (2020) and show that dropping gender would make the treatment policy more equitable. Overall, we highlight how the inclusion of group attributes can affect group fairness, to the extent of exacerbating inequalities in the data without necessarily affecting overall performance. This illustrates the theoretical arguments of Chapter 5, that approximating true probabilities can obscure the tangible effects of individual modelling choices. Moreover, it demonstrates that theoretical trade-off results do not preclude the existence of less discriminative alternatives in practice.

2.4.3 *Key Contributions*

- We derive tight upper bounds on the disagreement between classifiers of similar accuracy, and connect them to differences in disparate impact.
- For two model classes and three large datasets, each with two possible groupings, we demonstrate empirically that omitting group attributes can lead to much fairer classification rates without significant loss in accuracy.
- We show mathematically that under common conditions, calibrated predictors can exacerbate inequalities in the data, as measured by base rate difference and disparate impact.
- We show theoretically that logistic regression models using group attributes are particularly liable to this and demonstrate this phenomenon empirically.
- For a real-world labour market setting, we show that removing the gender attribute improves equity without harming performance, underscoring

that modelling choices need to be justified beyond claims to 'follow the real chances [...] as far as possible' (Allhutter et al., 2020).

2.5 CONSTRUCTING PEOPLE

This section motivates and summarises Chapter 7, 'Position: The Categorization of Race in ML is a Flawed Premise' (Doh et al., 2025).

2.5.1 *Personal Motivation*

Already in the early days of my PhD, I was sceptical of the reliance on race variables in testing for discrimination, as reflected in Chapter 3. This was influenced by Bob's thoughts on individuals and groups as well as by critical work such as Hu and Kohler-Hausmann (2020). During my research visit in Alicante, my friends and colleagues Miriam and Piera were planning a project to criticise US-centred race taxonomies in Europe from a mixed-race perspective, and started conducting interviews with community members. After I joined the project, it developed into a more general critique of the use of race variables in machine learning and, in particular, algorithmic fairness. Our points partly echo earlier critiques, such as by Moss (2021), but focus more on the discrepancy between current research practice and the lived reality and legal setting in Europe, in particular of individuals identifying as mixed-race. As also discussed by Selbst et al. (2019) and Birhane et al. (2022b), a fundamental issue with the algorithmic fairness literature is that, its abstract discussions—typical for computer science research—risk losing touch with lived experience. In our paper, we question common abstractions that over-simplify racial discrimination by reducing it to membership in groups that are assumed to be fixed.[7]

As Hacking (1986) argues, the tendency to put people into fixed categories is tied to the origins of statistics and to the purpose of control.[8] As for statistics, it is neither a coincidence that the 'scientific' investigation of racial differences was pushed by early statisticians (see Chapter 1), nor that the idea of 'natural kinds' originated in the seminal work of John Venn (1866) on the foundations of probability and statistics. As for control, the use of (different) race taxonomies for segregation, for example in the U.S. and South Africa, is well documented

7 In ongoing work, we aim to further bridge the gap to lived reality by organising workshops to connect research efforts with the perspectives of affected communities. Our first workshop is confirmed at CHI 2026: https://sites.google.com/view/between-and-beyond

8 This tendency 'one of the primary impulses of the modern state' according to Recht (2025b), see also Chapter 9.

(Bowker and Star, 1999). As these authors emphasise, categorisation is not necessarily bad, and often actually helpful. This does not conflict, however, with the imperative to be cautious with the categories we use, especially for sorting humans. We demonstrate, in line with Hacking (1986), that the act of categorisation is performative and can have grave downstream effects on how people are viewed. In Hacking's words, through this act, we are *making up people*.

2.5.2 *Summary*

In Chapter 7, which was published as a position paper, we point out problems with the race taxonomies that are commonly used in current machine learning research, particularly on fairness. We argue that the taxonomies are mostly U.S.-centric and do not generalise globally; we particularly discuss the European context—the case we are most familiar with—where racial discrimination operates along different dimensions and racial categories are not commonly used for administrative purposes. We also highlight the often-overlooked case of mixed-race identities, which do not fit these taxonomies and are consequently erased. We demonstrate that existing taxonomies cannot be adjusted to effectively consider mixed-race experiences; this is not a small design flaw, but highlights how fixed race taxonomies generally oversimplify lived reality. We also argue that the categorisation of race in digital labelling contributes to stereotyping and reification in general discourse, rolling back on decades of scholarship and policies trying to mitigate scientifically outdated race essentialism and pigeon-holing. We demonstrate empirically how such stereotypes surface in machine learning systems.

As an alternative, we advocate for developing alternative approaches that can better account for the complexity and contextuality of discrimination. For considerations in computer vision like dataset diversity, we support recent approaches focusing on phenotypes themselves rather than enforcing diversity of race labels based on a race classifier (Yucer et al., 2024a). We emphasise that we do not endorse a racial interpretation of these phenotypes, which would constitute a regression into scientifically obscure craniology (Crawford and Paglen, 2021). For tabular data, we advocate for the development of multidimensional approaches, which can take into account the interplay of multiple features that are relevant for discrimination in a given context. For example, a recent discrimination study in the Swiss labour market combined language, nationality, and name-based classification into an ethnicity denotation (Hangartner, Kopp and Siegenthaler, 2021). In contrast to their simplistic binary comparison of two resulting groups, we envision multi-dimensional analysis akin to ideas in indi-

vidual fairness (Dwork et al., 2012) or flexible notions of calibration as explored in Chapter 3. Overall, we identify important shortcomings of current research practice relying on simplistic and geographically limited race taxonomies: We argue that these are not sufficient for detecting racial discrimination and, on the contrary, can contribute to harmful stereotypes.

2.5.3 *Key Contributions*

- We demonstrate that fixed race taxonomies as used in algorithmic fairness research fundamentally fail to account for mixed-race identities, and argue that this failure is not a marginal design flaw but reveals a deeper incompatibility between rigid categorisation and lived racialised experience.
- We contrast the role and meaning of race categories in U.S. and European contexts, showing that assumptions implicit in U.S.-centred machine learning research do not carry over to Europe, where racial classification is legally restricted, socially contested, and operates through different proxies.
- We show that the digital reification of race labels in machine learning systems contributes to stereotyping and essentialism, counteracting longstanding scholarly and policy efforts to move beyond biologically grounded notions of race.
- We outline alternative directions for machine learning research that better respect the contextual and multi-dimensional nature of discrimination, including phenotype-based approaches in computer vision and multi-dimensional, context-sensitive analyses for tabular data.

2.6 CONSTRUCTING CAUSAL MODELS

This section motivates and summarises Chapter 8, 'The Accountability Gap in Causal Modelling for Automated Decision Systems' (Höltgen and Williamson, 2026a).

2.6.1 *Personal Motivation*

In the first year of my PhD, I realised that automated decision systems (ADSs) used for allocation problems need to model the outcomes under different treatments (or treatment and control) separately. To accurately predict the average

outcome of a given treatment rule, it is sufficient—and, realistically speaking, necessary—to be calibrated (in the general sense of (2.1)) in both predicting the average outcome under treatment of the treated and the average outcome under control for the untreated. Conversations with Konstantin Genin and Sebastian Zezulka then surfaced close connections to the potential outcomes framework by Neyman and Rubin. Concurrently, the literature on ADSs has witnessed a push towards causal modelling in recent years, criticising the limits of pure predictions that do not account for the decisions they inform (Liu et al., 2025; Wang et al., 2024; Zezulka and Genin, 2024).

For making consequential decisions about people, the dependence on untestable assumptions and predictions constitutes a lack of accountability: It gives decision makers more leeway to justify whatever policy they decide on, even more so than in the case of pure predictions (Chapter 5). This is due to the common dependence of causal models on untestable conditional independence assumptions and the dependence of individual decisions on 'estimated' individual causal effects—which are even more difficult to evaluate than probabilities (given that they are essentially differences between two probabilities). Beyond accountability, it is also difficult for the modellers themselves to evaluate the adequacy of their models, making it very hard to learn how well different models are doing on certain tasks (Vafa, 2024). I set out to develop a theoretical framework that distinguishes between treatment and control but does not rely on assuming the existence of data-generating distributions—as in Chapter 5—or unobservable but supposedly well-defined counterfactuals and individual causal effects.[9]

2.6.2 *Summary*

In Chapter 8, we draw attention to a fundamental tension in causal modelling for ADSs, between the aspiration to directly inform real-world decisions and the abstract nature of the available causal inference frameworks. We argue that Rubin Causal Models (RCMs) do not provide a satisfactory basis for accountability because they depend on untestable assumptions and do not make testable predictions. Drawing on literature across fields, we highlight that these assumptions entail strong stability and modularity requirements on the social world that are often not backed up by empirical investigations. In response, we propose an amended version of the framework that construes causal infer-

9 A particularly satisfying part of this work was showing how existing estimators fit into the proposed framework adaptation based on finite populations; while some estimators needed explicit targeting, while others emerged naturally from the framework and were basically 'reinvented'.

ence as the task of making treatment-wise predictions for finite target populations. Rather than formulating assumptions in terms of abstract distributions, our framework formulates assumptions as relationships between explicitly represented training and target populations, such that they can be tested in hindsight. This reframing shifts the focus from identifying true causal parameters to predicting properties of the (empirical) outcome distribution, such as its mean. We also formally connect the new assumptions to conventional ones, such as i.i.d. sampling and unconfoundedness, showing how the latter can be recovered as limiting or sufficient cases. We demonstrate that our predictive, population-based framework captures established causal estimators, such as matching, doubly robust estimators, and others, thereby shedding new light on them. By dissolving the sharp separation between internal and external validity, the framework forces explicit engagement with generalisation to the target population, an aspect that is required for decision-making yet often neglected. Overall, the paper argues that reframing causal inference towards prediction for concrete populations enables more transparent evaluation, clearer diagnosis of modelling failures, and, thus, a more solid foundation for accountability in ADSs.

2.6.3 *Key Contributions*

- We show that standard causal modelling impedes accountability of ADSs because assumptions and predictions cannot be directly evaluated.
- We draw on existing scholarship across fields to make the case that the assumptions commonly made in RCMs are often too strong in social settings.
- We present an adaptation of RCMs that directly models the relevant populations, makes testable predictions, and only relies on assumptions that can be checked in retrospect.
- We make the case that causal modelling through ADSs should be understood more modestly as making aggregate predictions rather than estimating individualised effects.
- We demonstrate that popular causal inference estimators can be analysed in the new framework, contributing to our understanding of them.

2.7 CONSTRUCTING INTELLIGENCE

This section motivates and summarises Chapter 9, 'A Thoughtful Narrative Emerges' (Höltgen, 2025).

2.7.1 *Personal Motivation*

In 2025, the editor-in-chief of the *Harvard Data Science Review*, Xiao-Li Meng, asked me if I would be interested in writing a peer-reviewed commentary for Sabina Leonelli's paper 'Environmental intelligence: Redefining the philosophical premises of AI' (Leonelli, 2025). I was enthusiastic not only because I was honoured to be invited to write a commentary for such a journal on the work of a researcher that I so deeply respect. I also welcomed the opportunity to follow and write up some thoughts on connections of my work to the general understanding of AI—this time not only for automated decision systems, but more broadly, including recent developments in Generative AI. In my commentary, I take Sabina's foundational critique of predominant AI narratives as a starting point to argue that common perceptions about the universality of current approaches to constructing intelligent systems overlook that the domains where AI systems excel are precisely those that fit more into the theoretical conception of machine learning. As I demonstrate throughout this thesis, this theory does not straightforwardly apply to the more messy social world, where the significance of abstraction steps and stability assumptions are often underestimated. Lastly, I took the commentary as an opportunity to connect my ideas about AI systems in society to a broader agenda on how to help shape a positive future. As many have pointed out before me (e.g. Kalluri, 2020; Kasy and Abebe, 2021), one of the most important questions about AI is how it affects power structures and how we can empower civil society rather than leaving the future of technology and society in the hands of a few.

2.7.2 *Summary*

In Chapter 9, I comment on Sabina Leonelli's vision of Environmental Intelligence, positioning it as a thoughtful alternative to the standard narrative which increasingly conflates intelligence with AI capabilities. To sharpen the profile of her ideas, I contrast them with two other recently articulated visions for rebranding AI as a technology in the service of humankind: the 'scientist AI' of Bengio et al. (2025) and the ideas of Acemoglu and Johnson (2023) around 'ma-

chine usefulness'. While Leonelli also places more emphasis on environmental considerations, my commentary focuses on the transdisciplinary epistemic considerations that stand out in her account. On the one hand, Leonelli draws out what I call the 'sterility' of common AI visions. Given that AI excels at well-defined tasks with pre-determined levels of abstractions (as discussed in Chapter 5), I argue that recent improvements in games or mathematical capabilities do not imply that AI systems have overcome the limits of situatedness and dependence on human judgement that Leonelli (2025) identifies. Reflecting the discussions of other chapters in this thesis, I then argue that the lack of 'unambiguous ground truth', which is sometimes considered an exception for ML tasks (Ziegler et al., 2020), is actually very common in the social world. This is not commonly acknowledged in the literature on machine learning and AI, but becomes quite clear in Leonelli's piece, in particular her discussion of the dependence on human judgement. Lastly, I emphasise the political and epistemic role of narratives in AI research, drawing on Acemoglu and Johnson (2023): How intelligence is framed shapes not only research priorities but also the socio-technological and economic trajectories that follow from them. Overall, my commentary supports Leonelli's narrative as a useful corrective that aligns more closely with the actual behaviour of current AI systems and with normative concerns about responsibility, power, and environmental impact.

2.7.3 *Key Contributions*

- I situate Leonelli's concept of environmental intelligence within contemporary AI discourse by contrasting it with other prominent reformulations of AI's societal role.
- I show that common AI narratives rely on assumptions of well-defined tasks, stable abstractions, and unambiguous ground truths, and argue that these are not as common as often assumed.
- I highlight the role of narratives in shaping AI research agendas and downstream socio-technological trajectories, as well as values in society more broadly. In line with Leonelli, I argue that the currently dominant narrative needs to be reconsidered.

Part II

PROBABILITY AND CALIBRATION

3

ON THE RICHNESS OF CALIBRATION

Legends of prediction are common throughout the whole House hold of Man. God speaks, spirits speak, computers speak. Oracular ambiguity or statistical probability provides loopholes, and discrepancies are expunged by Faith.

— Ursula K. Le Guin, *The Left Hand of Darkness*

AUTHOR CONTRIBUTIONS

The idea and conceptualisation of the project are mostly due to Benedikt Höltgen, based on many discussions with Robert Williamson. Theoretical results and writing are due to Benedikt Höltgen, with valuable feedback by Robert Williamson.

ABSTRACT

Probabilistic predictions can be evaluated through comparisons with observed label frequencies, that is, through the lens of calibration. Recent scholarship on algorithmic fairness has started to look at a growing variety of calibration-based objectives under the name of multi-calibration but has still remained fairly restricted. In this paper, we explore and analyse forms of evaluation through calibration by making explicit the choices involved in designing calibration scores. We organise these into three grouping choices and a choice concerning the agglomeration of group errors. This provides a framework for comparing previously proposed calibration scores and helps to formulate novel ones with desirable mathematical properties. In particular, we explore the possibility of grouping datapoints based on their input features rather than on predictions and formally demonstrate advantages of such approaches. We also characterise the space of suitable agglomeration functions for group errors, generalising previously proposed calibration scores. Complementary to such population-level scores, we explore calibration scores at the individual level and analyse their

relationship to choices of grouping. We draw on these insights to introduce and axiomatise fairness deviation measures for population-level scores. We demonstrate that with appropriate choices of grouping, these novel global fairness scores can provide notions of (sub-)group or individual fairness.

3.1 INTRODUCTION

Enabled by the now widespread availability of data and compute, many data-driven tasks are increasingly aided by machine learning systems. Examples of this include the assessment of individual risk, for example for credit default. This comes with its own risks such as structurally biased data collection (O'Neil, 2017), reinforcement of power structures (Kasy and Abebe, 2021) or algorithmic monoculture (Kleinberg and Raghavan, 2021; O'Neil, 2017). It does, however, have the advantage that the deployed algorithms can be thoroughly assessed, also with regard to notions of fairness. As Vredenburgh (2022) argues, fairness is a standard of justice (among others) that can be defined as the requirement of 'respecting people's claims in proportion to the strength of those claims'. In this work, we concentrate on fair probabilistic predictions *in light of given data* – in particular, on notions of fairness based on calibration. While this is somewhat restrictive, it is – contrary to more 'substantive' notions of fairness (Green, 2022) – something formal approaches are well suited for; while there is a place for formal work on fairness, it cannot stand alone (Selbst et al., 2019).

The term 'calibration' can refer to either the agreement of predictions with the frequency of observed labels or the process of aligning them; we are concerned with and use it in the former sense. Vredenburgh (2022) notes that '[m]any computer scientists, social scientists, and philosophers favour calibration as the best metric for fairness'. To sensibly compare predictions with label frequencies, it is necessary to group datapoints together. This is most commonly done via bins of the prediction space $[0, 1]$, with some variations especially in recent years – sometimes in the context of fairness and then often under the name of multi-calibration (Hébert-Johnson et al., 2018). A step in the direction of local notions of calibration has been made in (Luo et al., 2022) and (Cabitza, Campagner and Famiglini, 2022), each proposing a global score with components that can be localised in a learned representation space and in prediction space, respectively. An ethical analysis of calibration-based notions of fairness requires an understanding of the choices involved in arriving at such notions. However, these choices have hardly been explicitly discussed or analysed in previous literature. In this work, we present a conceptual framework from the ground up with the goal of generalising previous notions of (multi-)calibration and making relationships

between them more explicit. Contrary to the understanding of calibration on subgroups as a *refinement* of group calibration, we demonstrate that calibration originates at a localised level and is then extended to groups.

We elaborate on the emergence of groups as a fundamental aspect of calibration in practice and on the historical development of its now prevalent understanding in Section 3.2. A major goal of this work is then to point to the vast space of possible alternatives and analyse the choices involved. We demonstrate advantages and disadvantages of such alternatives for assessing prediction quality and fairness. Where possible, we capture salient aspects of these trade-offs mathematically. We show that calibration allows for both global and local evaluation, that is, both at the population level and localised to some point in input space. The former involves the additional choice of an agglomeration function over group calibration errors. The choices involved in global calibration scores as well as the structure of the present paper are captured in the following diagram:

$$\underbrace{\text{principle} \quad \text{size} \quad \text{structure}}_{\text{grouping}} \quad \text{agglomeration}$$

Most fundamental is the choice of a general *principle* according to which datapoints are grouped together. This can be arbitrary but will usually be some notion of vicinity in input space or prediction space (Section 3.3.1). Independent of the grouping principle, there is a trade-off concerning the *size* of the group between robustness and informativeness (Section 3.3.2). The third choice involved in the selection of a grouping concerns its general *structure*. Here, the options include bins of fixed size but also overlapping groups of nearest neighbours or kernels (Section 3.3.3). The remaining choice involved in a global score is the *agglomeration* function applied to the group errors. We show that the commonly used maximum and average of absolute group errors are the two extreme cases of a broader class of functions suitable for evaluating prediction quality (Section 3.4). To introduce population-level fairness scores based on calibration, we then axiomatise a matching class of deviation measures (Section 3.5.1). Our analysis of the richness of possible choices culminates in a demonstration that calibration errors on suitable groups, agglomerated by a deviation measure, can provide scores that capture notions of either individual or (sub-)group fairness (Section 3.5.2). We summarise our contributions as follows:

- We show that calibration can be construed more broadly than often assumed, which allows novel approaches to evaluate probabilistic predictions.

- We organise the space of possible calibration scores through a small number of largely independent choices, pointing out considerations and trade-offs relevant to these choices. We capture salient aspects of these trade-offs mathematically.
- To our knowledge, we propose the first provably consistent estimators of individual calibration, based on kernels and a novel notion of k-nearest neighbour-based groupings.
- We characterise the space of functions suitable for agglomerating calibration errors into global scores of prediction quality, capturing existing notions of calibration.
- We propose and axiomatise agglomeration functions for global calibration-based fairness scores and show that through suitable choices of grouping, such scores can provide notions of (sub-)group fairness as well as notions of individual fairness.

3.2 CALIBRATION FROM THE GROUND UP

A common definition of (perfect) calibration is that for all $v \in [0,1]$, among the datapoints which receive the prediction v, the (expected or observed) frequency of the label '1' is equal to v. In this section, we motivate our perspective on calibration as a notion that actually originates at the individual level and requires choices of grouping in practice. We also provide a tentative historical explanation for why the understanding of calibration is usually more narrow. Throughout the paper, we investigate the problem of predicting binary labels $\{0,1\}$ from inputs in a space $\mathcal{X}$.

3.2.1 *From individuals to groups*

Given access to a dataset $\{(x_i, y_i)\}_{i\in\mathbb{N}}$ with infinite datapoints for each input $x \in \mathcal{X}$,[1] a probabilistic predictor $p : \mathcal{X} \to [0,1]$ of binary labels would be 'correct' if it

1 Within this particular paragraph, and only here, we assume $\mathcal{X}$ to be countable for ease of presentation.

predicts the true probability everywhere, as reflected in the limiting frequency of the label '1'.[2] Thus, a probabilistic predictor p should satisfy

$$\forall x \in \mathcal{X}: \quad p(x) = \lim_{n\to\infty} \frac{1}{|V_x^n|} \sum_{i\in V_x^n} y_i \quad \text{with} \quad V_x^n := \{i \leqslant n : x_i = x\}. \tag{3.1}$$

In practice, of course, we do not have access to infinite data.

From now on, we assume a finite dataset $\mathcal{D} := \{(x_i, y_i)\}_{i\in\mathcal{I}}$ with $\mathcal{I} := \{1, ..., N\}$ and $\mathcal{X}_\mathcal{D} := \{x_i\}_{i\in\mathcal{I}} \subset \mathcal{X}$. If we had lots of data for every input, we could still impose a criterion of perfect point-wise, 'individual' calibration. Now based on limited data, this would take the similar form

$$\forall x \in \mathcal{X}_\mathcal{D}: \quad p(x) = \frac{1}{|V_x|} \sum_{i\in V_x} y_i \quad \text{with} \quad V_x := \{i \in \mathcal{I} : x_i = x\}. \tag{3.2}$$

The availability of such abundant data would arguably render the deployment of machine learning models superfluous and suggest that the representation of individuals in the data is too coarse to treat them adequately.

In interesting settings, we have at most a few datapoints per input, and usually no data at all for some inputs. This is, of course, why we need machine learning algorithms with constrained hypothesis classes and inductive biases in the first place. To compare label frequencies with predictions 'at x', we then need to somehow extend the considered groups from those consisting only of identical inputs V_x to more general groups of datapoints $I_x \subset \mathcal{I}$.[3] This then leads to the more general criterion of calibration

$$\forall x \in \mathcal{X}_\mathcal{D}: \quad \frac{1}{|I_x|} \sum_{i\in I_x} p(x_i) = \frac{1}{|I_x|} \sum_{i\in I_x} y_i. \tag{3.3}$$

Note that, in contrast to (3.1) and (3.2), this does not suggest an 'optimal' value of $p(x)$. Instead, it provides a quality criterion that can be localised to a particular point in input space. As Dawid notes, '[t]he calibration criterion has some similarity with the frequency definition of probability, but does not require a background of repeated trials under constant conditions' (Dawid, 1982, p. 606). Based on this, we can define the (signed, i.e. positive or negative) *calibration error* of a group $I \subset \mathcal{I}$ to be

$$c(I) := \frac{1}{|I|} \sum_{i\in I} (p(x_i) - y_i). \tag{3.4}$$

2 Note that in general, such a limit need not exist. While the law of large numbers is often taken to show that such limits exist and do reflect the true probability with probability one, this depends on the particular measure imposed on the infinite-dimensional space of possible sequences (Schnorr, 2007, p. 8). We generally do not assume the existence of a true probability distribution in this paper, except where explicitly stated.

3 To avoid introducing additional noise, it is desirable to always take *all* datapoints that instantiate a given input, i.e. $I_x = \bigcup_{z\in U_x} V_z$ for some suitable subset of the input space of $U_x \subset \mathcal{X}$.

3.2.2 *Conventional understanding of calibration*

A lot of work on calibration can be seen as defining the I_x via pre-defined bins $B_1, ..., B_K$ forming a partition of $[0, 1]$ (Zadrozny and Elkan, 2001). The most common example of this is the Expected Calibration Error (ECE; Naeini, Cooper and Hauskrecht, 2015) defined as

$$\mathsf{ECE}(p) = \sum_{k=1}^{K} \frac{|B_k|}{N} \left| \sum_{i:p(x_i)\in B_k} p(x_i) - y_i \right|. \tag{3.5}$$

While often assumed to be *the* way to think about calibration, such prediction-based grouping is only one very specific choice, and not always suitable. It does have the advantage that it allows for quickly assessing general under- or over-confidence of predictors, a question that has proven to be particularly relevant for modern neural networks (Guo et al., 2017). For a formulation of the ECE (and other calibration scores) more in the style of (3.3), see Appendix 3.7.1.

However, there also seem to be contingent historical reasons for the prevalence of prediction-based grouping for calibration. The evaluation of probabilistic predictors was primarily researched in the meteorological community and published in meteorology journals until well into the 70s, the Brier score (Brier, 1950) being perhaps the first notable work. Here, predictions were made by human forecasters – potentially with the help of automated procedures (Sanders, 1963) – and their evaluations only had access to predictions and observed outcomes, not to the inputs.[4] Furthermore, these forecasters only assigned probabilities on a finite scale, such as multiples of 10%, which did not require to impose bins and, thus, made these groups seem even more 'natural'. Calibration in the sense of (3.1) - (3.3) was often called 'reliability' or 'validity', while 'calibration' was used in relation to over-and under-confidence (Sanders, 1963), closer to the original meaning of the word (regarding the grounding of scales or instruments).

When statisticians subsequently picked up the topic (and seem to have renamed it 'calibration'), they continued to use weather forecasting as the primary example and modelled the setup as a sequence of forecasts and outcomes. They also extended the notion of calibration to basically the general setting of (3.3), speaking of subsequences rather than groups: Dawid (1982) admits roughly any subsequence that does not take into account the ground-truth labels. In more

4 Interestingly, it seems that these forecasters used a strategy of '*sorting* all instances into an ordered set of categories of likelihood of occurrence, and *labelling* each category with a specific likelihood, or probability, of occurrence' (Sanders, 1963, p. 200, original emphasis). This fits nicely with then assessing the prediction quality through calibration based on groups of similar *prediction*.

recent work, Dawid (2017) suggests that also the information on which predictions are based can be taken into account for '$\mathcal{H}$-based' grouping, but the discussion remains on an abstract level.

Roughly concurrently, work on 'multi-calibration' has suggested explicitly taking into account input information for selecting groups: The original paper on multi-calibration (Hébert-Johnson et al., 2018) proposed to look at all computable groups – a proposal that has also been made in the sequence setting by (Dawid, 1985). But even approaches based on similarity of learned representations either *prescribe* that the selected groups have equal predictions (Burhanpurkar et al., 2021) or apply additional binning in prediction space (Luo et al., 2022).[5] Approaches proposing to omit this additional binning are sometimes referred to as 'multi-accuracy' (Hébert-Johnson et al., 2018; Kim, Ghorbani and Zou, 2019) but we are not aware of a thorough discussion of this choice. We find this name confusing since a predictor could be very wrong on every datapoint but still 'multi-accurate' – this seems to not be in line with established notions of accuracy.[6] By taking into account the available structure in the input space, different notions of vicinity can be used to construct groups I_x 'around x'. Calibration errors of such groups allow to recover a localised sense of calibration – we return to this in sections 3.3.3 and 3.5.2. In the following section, we discuss considerations concerning the choice of grouping.

Intuition (Section 3.2): To assess a predictor p at a point x, we could compare it to the average label at x if we had an infinite (3.1) or a large (3.2) number of datapoints for x. Since we usually do not have many datapoints for a given x, we can instead take a group of datapoints I_x including it and compare their *average* prediction to their ratio of positive labels (3.3). For partly historical reasons, I_x has almost exclusively been taken to be the group of datapoints that is similar to x in their *predictions*.

3.3 GROUPING CHOICES

Rather than a requirement, assessing calibration on prediction-based bins is only one particular choice. This naturally leads to the question of what other options there are to group datapoints. In the following, we constrain ourselves to groupings based on vicinity in the input- or prediction space, that is, based on

5 A very recent exception is (Kelly and Smyth, 2023), suggesting to only use ECE-style binning on a single input feature like age.

6 Consider a setting where in each group, there are equally many 1s as 0s among the labels and the predictor always predicts $p(x_i) = 1 - y_i$ or always predicts 0.5. In both cases, the average label per group, 0.5, coincides with the average prediction in every group.

notions of cohesion and similarity. One alternative, suggested in (Burhanpurkar et al., 2021), would be to consider all efficiently computable groups. Vicinity-based groups already give a wide space to explore and make groups and their errors particularly interpretable, which is helpful both for improving predictors and for questions of fairness. Note that one may also pursue different approaches simultaneously to gain more insights, given that there is no 'correct' way of grouping. In the words of Nelson Goodman, 'similarity is relative and variable, as undependable as indispensable' (Goodman, 1972, p. 20).

To better understand a large space of possible choices, it is often helpful to consider inherent trade-offs. For the case of grouping datapoints, these include vicinity in the input space versus in prediction space (Section 3.3.1) as well as the size of groups (Section 3.3.2). The third choice concerns the grouping structure, that is, how a notion of vicinity is applied to form groups (Section 3.3.3).

3.3.1 *Vicinity-based grouping principles*

Notions of vicinity particularly relevant to grouping datapoints are those of group cohesion and individual similarity. We formalise cohesion as a property of topological spaces, namely connectedness. Note that we review basic notions of general topology in Appendix 3.7.2. We say that a grouping is similarity-based if it makes use of distances between points, e.g. through a metric. In the sense that every metric space induces a topological space and every ball is a connected set in the induced topology but not vice versa, cohesion is weaker than similarity. Cohesion is usually not sufficient to define groups, which means that vicinity-based groups generally rely on some notion of similarity.

A choice inherent in building groups based on vicinity is whether to focus on predictions $p(x_i)$ (as in the ECE) or input features x_i.[7] As described in the previous section, the popularity of prediction-based grouping may be partly due to the roots of calibration in meteorology where no useful structure in the input space was available for evaluating predictors. Its main advantage is that it facilitates the assessment of over- and under-confidence (in terms of distance from 0.5), by separating datapoints with high predictions from those with low predictions. On the other hand, grouping datapoints based on vicinity in $\mathcal{X}$ allows evaluating p in specific areas of the input space. This can be desirable, for example, in the context of fairness where we are interested in the performance of p on groups that are often in some way local or similar in the input space. Some

7 Interestingly, similarity of input and of prediction can, respectively, be approximately mapped to what Campbell (1958) identified as the two conditions for 'entitativity' of groups: similarity and common fate.

recent approaches propose to group points based on similarity of representations learned by some feature extractor (Burhanpurkar et al., 2021; Luo et al., 2022) An advantage of such approaches is that they implicitly make use of both inputs and predictions while a disadvantage is that they do so in an opaque way. Especially for fairness applications, this might be difficult to justify since the ethical significance of such groups is less clear compared to groups defined explicitly in input or prediction space.

One important aspect of choosing a notion of vicinity is the mathematical structure that needs to be assumed. Approaches to group fairness and multicalibration mostly assume some set structure, often a partition given by individual categorical features (e.g. race intersecting with gender). Relying on $\mathcal{X}$-similarity requires that a sensible notion of distance such as a metric is available in the input space.[8] However, we are not usually given a metric that captures similarity in a principled way. One option would then be to start with a standard metric (e.g. based on $\|\cdot\|_p$) and weight each feature dimension based on its observed variance or range, in an attempt to give each dimension equal relevance for the assessment of similarity. Note that the exact form of the metric is not as central here as for Lipschitz-based notions of individual fairness; however, although it here only indirectly influences calibration scores via the grouping, it still remains a concern. In this sense, $\mathcal{X}$-similarity requires more additional structure than prediction similarity since the interval $[0,1]$ comes with much less controversy about the choice of metric: All $\|\cdot\|_p$-based metrics induce the same intervals in $[0,1]$ and there is no issue of balancing multiple dimensions. While there might be contexts where e.g. distances between points close to 0 or 1 should be elevated as compared to points close to 0.5, the metric $|v - v'|$ seems generally well-suited to capture prediction similarity. Also note that groupings based on learned representation implicitly use structure in both $\mathcal{X}$ and $[0,1]$, through inductive biases of the extractor.

A trade-off between notions of vicinity arises due to the fact that disconnected areas in the input space may be mapped to the same predictions (or, analogously, the same representations), leading to different notions coming apart. We characterise this tension in the language of topology. The proofs for all propositions can be found in Appendix 3.7.3.

Proposition 3.1 (Cohesion in $[0,1]$ and in $\mathcal{X}$)**.**

a) For continuous $p : \mathcal{X} \to [0,1]$*, connectedness in* $\mathcal{X}$ *implies connectedness in* $[0,1]$ *but not vice versa.*

8 Strictly speaking, k-nearest neighbour approaches only require a complete and transitive ordering for every datapoint.

b) For any non-constant function $p : \mathcal{X} \to [0,1]$*, There is a topology on* $\mathcal{X}$ *s.t.* p *is not continuous (and thus connectedness in* $\mathcal{X}$ *does not imply connectedness in* $[0,1]$*).*

c) If p *has more than a single local minimum or more than a single local maximum, there are intervals in* $[0,1]$ *whose preimage in* $\mathcal{X}$ *is not connected in any Hausdorff topology on* $\mathcal{X}$*.*

Groups of interest for fairness considerations usually depend on input features rather than predictions, such as groupings along lines of age, gender, race, and/or socio-economic background. Proposition 3.1 shows that unless p is very restricted, prediction-based grouping can lead to common assessment (in terms of a single calibration error) of areas in $\mathcal{X}$ that are not connected. A potentially problematic case in the context of fairness is when a small minority group (assumed to be similar to each other in $\mathcal{X}$) have predictions similar to another, larger group such that the predictor's behaviour on the small group is washed out in the error of the combined group. Another problem, not directly related to topology, is that of 'stereotyping', i.e. when predictions are similar in two (potentially adjacent) areas of $\mathcal{X}$ in such a way that p over-predicts in one but under-predicts in the other: this may suggest perfect calibration in the combined group as its average prediction may still coincide with the average label (cf. Proposition 3.2). $\mathcal{X}$-based grouping can detect such stereotyping while prediction-based grouping cannot. Such problems are the reason why scholars have proposed to test calibration in multiple subgroups – however, this still allows for such effects within these subgroups.

3.3.2 *Group size*

The trade-off between vicinity in input space and in prediction space cannot be resolved by arbitrarily intersecting $\mathcal{X}$-similar with p-similar groups. This is one implication of the more obvious trade-off concerning the size of the groups: Larger groups allow more stable estimates while errors of smaller groups contain more information about their members.[9] We formalise and quantify this in the following.

Proposition 3.2 (Advantage of smaller groups: More information)**.**
Take two nonempty, disjoint sets of indices $I_1, I_2 \subset \mathcal{I} = \{1, ..., N\}$ *and recall the definition of the calibration error in (3.4). Then*

9 Dawid (2017) mentions a related but slightly different trade-off between robustness and incisiveness w.r.t. the amount of information taken into account for groups on which to calibrate forecasts.

a) Assume that $\forall(i,j) \in I_1 \times I_2 : x_i \neq x_j$ *and that not all labels* $y_i, i \in I_1 \cup I_2$ *are equal. Then a predictor* $p : \mathcal{X} \to [0,1]$ *can be perfectly calibrated on* $I_1 \cup I_2$ *but neither on* I_1 *nor on* I_2.

b) For any predictor $p : \mathcal{X} \to [0,1]$, $|c(I_1)| \leqslant \alpha$ *and* $|c(I_2)| \leqslant \beta$ *entails* $|c(I_1 \cup I_2)| \leqslant \frac{\alpha \cdot |I_1| + \beta \cdot |I_2|}{|I_1| + |I_2|}$. *For* $\alpha = \beta$, *this means* $|c(I_1 \cup I_2)| \leqslant \alpha$.

Proposition 3.3 (Advantage of larger groups: Variance)**.**
Let $(X_1, Y_1), ..., (X_K, Y_K)$ *be pairs of i.i.d. random variables with values in* $\mathcal{X} \times \{0,1\}$ *and joint distribution* P. *Take some* $p : \mathcal{X} \to [0,1]$ *and define the Bayes-optimal predictor* $p^\star(x) := \mathbb{E}_P[Y|X = x]$. *Then*

$$\mathbb{V}_P\left[\frac{1}{K}\sum_{i=1}^{K} p(X_i) - Y_i\right] = \frac{1}{K}\mathbb{E}_P\left[p^\star(X_1)(1 - p^\star(X_1))\right] : \tag{3.6}$$

Proposition 3.2a) shows that finer groups or partitions can detect more problems with the predictor. Relating this to the above-discussed problem of stereotyping in the fairness context gives an argument for using finer groupings or partitions. However, while larger groups may overlook more forms of stereotyping, it is also not feasible to go arbitrarily small: Smaller groups come at the cost of giving more 'unstable' calibration errors in the sense that a group of K datapoints drawn independently from a true distribution P has an error variance that is inversely proportional to K (Proposition 3.3). Note that the forms of these results also highlight that the group size trade-off cannot be tackled in a general way as its character strongly depends on the distribution of data.

3.3.3 *Alternative grouping structures*

We now turn to the third choice involved in the grouping of datapoints, its general structure. As seen e.g. in the ECE (3.5), a popular way to construct groups based on a metric is by using bins. The necessary choice of number and size of bins is further complicated by the question of whether to make the individual bins equally large or to make them contain equally many datapoints (or some other scheme).[10] Still, bins are but one specific example of a partition and yet another question is whether to use a partition in the first place. They have the advantage of allowing a natural hierarchy of groupings since unions of non-intersecting groups are themselves sensible groups. One can also construct partitions based on input features, e.g. using a multi-dimensional grid instead of one-dimensional bins. We already discussed advantages of finer input-based

10 Some considerations in the case of prediction-based bins are discussed in (Nixon et al., 2019; Vaicenavicius et al., 2019).

partitions over partitions induced by prediction-based bins above. In the following, we discuss two alternative grouping structures, which can be localised in input space.

One option is to choose overlapping groups around individual points $x \in \mathcal{X}$ or $p(x) \in [0,1]$ by selecting their k nearest neighbours (k-NN) in the dataset according to some notion of distance. In contrast to fixed-size balls, k-NN directly controls the group size, independently of the distribution and size of the dataset. A downside in comparison to partitions is that some datapoints may be part of many more groups than others.

Proposition 3.4 (Group overlap for $\mathcal{X}$-based k-NN)**.**
Take $\mathcal{X} = \mathbb{R}^d$ and the grouping $\{I_j\}_{j\in\mathcal{J}}$ based on k-NN of datapoints $x_1, ..., x_N$ using any L^r-based metric on $\mathcal{X}$ with $r \in \mathbb{N}$. Depending on the dataset, it is possible that one datapoint is part of $\min(N,\ 2 \cdot d \cdot (k-1)+1)$ groups while another datapoint is part of only one group (unless $N \leqslant k$).

When group errors are agglomerated via a function satisfying Translation Equivariace for the purpose of evaluating prediction quality (Section 3.4), this may result in an unwanted disproportionate influence of some datapoints compared to others. However, $\mathcal{X}$-based k-NN not only provides errors that can be localised in input space, it can even be shown that its finite data estimate of the calibration errors converge to their 'true' value in the limit of infinite data (assuming the existence of a true distribution P) when $\mathcal{X} = \mathbb{R}^d$. In other words, such groups can provide consistent estimators of the true individual calibration error $p(x) - \mathbb{E}_P[Y|X = x]$ at each $x \in \mathcal{X}$. Note that the existence of a true probability distribution is a questionable assumption. Still, such theoretical results are always quite a positive sign.

Proposition 3.5 (Strong consistency of $\mathcal{X}$-based k-NN)**.**
Assume $\mathcal{X} = \mathbb{R}^d$, a joint distribution P on random variables (X, Y) with values in $\mathcal{X} \times \{0,1\}$ with Bayes-optimal predictor $p^\star(x) := \mathbb{E}[Y|X = x]$, a predictor $p : \mathcal{X} \to [0,1]$ and a metric $d_\mathcal{X}$ on $\mathcal{X}$ s.t. $(\mathcal{X}, d_\mathcal{X})$ is separable. Then

$$\mathbb{E}_P\left[\left|c(I_X^N) - (p(X) - p^\star(X))\right|\right] \to 0 \text{ with probability one as } N \to \infty, \tag{3.7}$$

where I_x^N denotes the set of k-NN of x with $k \to \infty$, $\frac{k}{N} \to 0$ for $N \to \infty$.[11]

The advantage and disadvantage of k-NN over partitions are also shared by kernel-based notions of calibration, to a degree determined by the chosen bandwidth. Often based on some metric, kernels can be used to weigh the influence

11 This slightly simplified formulation glosses over technicalities regarding the treatment of ties. For a comprehensive discussion, see (Devroye et al., 1994). Our proof reduces the problem to k-NN consistency of the variable $Z := p(X) - Y$ (see Appendix 3.7.3).

of datapoints in proportion to some notion of distance to x or $p(x)$. Luo et al. (2022) propose kernels in a learned representation space intersecting with prediction bins (see Appendix 3.7.1), thus interpolating between local calibration and the standard prediction binning via the bandwidth parameter. They also show that their estimator is consistent, but with respect to the true group error rather than the individual calibration error. To capture such kernel-based notions, we generalise calibration errors for discrete groups, defined in (3.4), to calibration errors for normalised distributions q on $\mathcal{X}$:

$$C(q) := \mathbb{E}_q[p(X) - p^\star(X)]. \tag{3.8}$$

It can easily be checked that we can express the calibration error $c(I_j)$ of a group I_j through $C(q_j)$ via the distribution

$$q_j(x) := \frac{|\{i \in I_j : x_i = x\}|}{|I_j|}, \tag{3.9}$$

thus generalising the notion of a group to a more abstract entity, a distribution on $\mathcal{X}$. Note that for this, we need to assume that groups always contain either all datapoints instantiating some $x \in \mathcal{X}$ or none – that is, as a union of groups V_x from (3.2), see also footnote 3. There is generally no reason to include only some datapoints instantiating a given input but not others, especially since equal input also implies equal prediction. This is corroborated by the observation that all groupings used in established calibration scores, surveyed in Appendix 3.7.1, satisfy this. We can also prove the consistency of $\mathcal{X}$-based kernels as estimators of the true calibration error:

Proposition 3.6 (Strong consistency of $\mathcal{X}$-based kernels)**.**
Assume $\mathcal{X} = \mathbb{R}^d$, a joint distribution P on random variables (X, Y) with values in $\mathcal{X} \times \{0, 1\}$ with Bayes-optimal predictor $p^\star(x) := \mathbb{E}[Y|X = x]$, and a predictor $p : \mathcal{X} \to [0, 1]$. Then

$$\mathbb{E}_P\left[\left|C(q_X^\gamma) - (p(X) - p^\star(X))\right|\right] \to 0 \text{ with probability one as } N \to \infty, \tag{3.10}$$

where $q_x^\gamma(x') := \frac{1}{\gamma}k(x - x') \cdot \frac{1}{K}$ with $K := \sum_{i=1}^N \frac{1}{\gamma}k(x - x_i)$ denotes the normalised distribution induced by a Borel kernel k with the bandwidth satisfying $\gamma \to 0$ and $\frac{N \cdot \gamma^d}{\log N} \to \infty$ for $N \to \infty$. See Appendix 3.7.3 for the constraints on k.

Note that the bandwidth parameter faces a similar trade-off to that of the group size but interpolates more smoothly. A general takeaway from this section should be that as for 'p is fair', the statement 'p is calibrated' is not a sensible statement as this is always relative to the chosen grouping. For calibration-based evaluation, a transparent communication of the grouping choices is therefore crucial.

Intuition (Section 3.3): The formation of groups that is necessary for evaluating calibration involves a number of choices. One concerns the notion of vicinity, in particular, whether groups should be based on vicinity in input space or in prediction space (Proposition 3.1). A clear trade-off concerns the size of the groups: smaller groups provide more detailed information (Proposition 3.2) while larger groups provide a more stable error estimation (Proposition 3.3). The third choice concerns the way a grouping is structured, which can be done via partitions or overlapping localised groups. Prime examples of the former are bins while the latter may use k-NN or kernels, both of which can provide consistent estimators of individual calibration scores (Propositions 3.5 and 3.6).

3.4 GLOBAL CALIBRATION SCORES

Thus far, we have only considered group-wise calibration errors. In order to evaluate predictors or to provide an objective that can be optimised, it is often helpful to condense an evaluation into a single score. In this section, we analyse the space of sensible agglomeration functions turning group errors into a global, population-level score. The motivation for exploring the space of sensible functions is threefold: First, situating established agglomeration functions in this space can provide a deeper understanding of their character. Second, it may inspire the use of alternative agglomeration functions from this space. Third, it allows connecting the grounding of global fairness scores (Section 3.5.1) in this more familiar setting, both implicitly by building intuition for agglomeration functions and explicitly by establishing mathematical connections.

In Appendix 3.7.1, we survey established calibration scores and show how our framework captures them, thereby allowing direct comparisons between them in terms of the choices discussed in this work. Two observations relating to this survey are that all the scores agglomerate *absolute* calibration errors $|c(I_j)|$ *via the average or the maximum*. In line with these and all other calibration scores that we are aware of, we only consider the agglomeration of absolute calibration errors in this section, thereby taking over- and under-prediction to be equally undesirable. In the following, we assume a set of absolute calibration errors indexed by a set J on which we impose a uniform probability measure. One could instead simply consider agglomeration functions on finite index sets, that is, on tuples of errors, but we decided to work in this more general setting that also allows working with a non-uniform measure, such as the probability P on

$\mathcal{X}$ in case $J = \mathcal{X}$. It also allows to leverage theoretical machinery applicable to uncountable sets.

Definition 3.7 (Agglomeration Function)**.**
For a measure space J*, an* agglomeration function *is a function* $R : \mathcal{L}^2(J) \to \mathbb{R}$.

$\mathcal{L}^2(J)$ denotes the space of real-valued random variables on the index set J with finite second moment; that is, elements of $\mathcal{L}^2(J)$ assign to every index $j \in J$ a real number, which should here be thought of as the absolute calibration error of the group I_J or q_j. We use the term 'agglomeration function' rather than 'aggregation function' since the latter is often defined in a more specific way (Grabisch et al., 2009).

We suggest that for any $C, C' \in \mathcal{L}^2(J)$ and any $\alpha \in \mathbb{R}$, an agglomeration function R should satisfy the following axioms:

A1	**Monotonicity**	$R(C) \leqslant R(C')$ if $C \leqslant C'$ almost surely.
A2	**Translation Equivariance**	$\forall \alpha \in \mathbb{R} : R(C + \alpha) = R(C) + \alpha$.
A3	**Positive Homogeneity**	$R(0) = 0$ and $\forall \mu > 0 : R(\mu C) = \mu R(C)$.
A4	**Subadditivity**	$R(C + C') \leqslant R(C) + R(C')$.
A5	**Law Invariance**	$R(C) = R(C')$ if C and C' have the same distribution of values.

A1 is the requirement that larger errors should lead to larger global scores. **A2** and **A3** are desirable properties for agglomerating group-wise calibration errors into global calibration scores as they preserve the naturally interpretable scale of the group-wise errors. While being less interpretable, scores that do not satisfy these axioms might still be useful as objective functions. **A4** is a particularly intuitive requirement when C and C' are the restrictions of some $D \in \mathcal{L}^2(J)$ to some partition $\{J_0, J_1\}$: The sum of the scores on J_0 and J_1 should not be smaller than the score of D. Also note that A4 is equivalent to Convexity[12] under A3 (Artzner et al., 1999) which makes R an easier optimisation target. **A5** ensures that all groups are treated equally and is sometimes also called Anonymity. Note that we may still weight groups according to their size by choosing a grouping that includes groups multiple times, as shown for the ECE in Appendix 3.7.1. It turns out that these axioms are central to the theory of risk measures, which allows to draw on insights from this rich literature.

Definition 3.8 (Coherent Risk Measure; Artzner et al., 1999)**.**
An agglomeration function satisfying A1-A4 is called coherent risk measure (CRM).

Definition 3.9 (Conditional Value at Risk; Artzner et al., 1999)**.**
Let $Q_C(\alpha) := \sup\{\mu \geqslant 0 : P(C \geqslant \mu) < \alpha\}$ *denote the quantile function for* $C \in \mathcal{L}^2(J)$,

12 Convexity is the condition that $\forall \lambda \in (0,1) : R((1-\lambda)C + \lambda C') \leqslant (1-\lambda)R(C) + \lambda R(C')$.

$\alpha \in [0,1]$. *The* conditional value at risk *is defined as* $\mathrm{CVaR}_\alpha(C) := \mathbb{E}[C \mid C > Q_C(\alpha)]$ *for* $\alpha \in [0,1]$.[13]

Any law-invariant CRM R – and thus any agglomeration function satisfying A1-A5 – can be built from CVaR_α via

$$\mathrm{R}(C) = \sup_{\nu \in \Lambda} \int_0^1 \mathrm{CVaR}_\alpha(C)\, \mathrm{d}\nu(\alpha) \tag{3.11}$$

for some set Λ of probability measures over $[0,1]$ (Kusuoka, 2001). Given that $\{\mathrm{CVaR}_\alpha\}_{\alpha\in[0,1]}$ interpolates between the expectation and the maximum (attained for $\alpha = 0$ and 1, respectively), law-invariant CRMs as a whole populate the space between expectation and maximum. In this sense, all previously proposed calibration scores are extreme cases of sensible calibration scores.

As we have divided the design of global calibration errors into grouping and agglomeration, it is interesting to ask whether one is strictly more powerful than the other, that is, whether choices for one of them can be 'simulated' by choices of the other. First, note that they take into account different kinds of information. The grouping looks at inputs and predictions without considering the resulting group errors while the agglomeration only looks at the resulting group errors. An example where the grouping already determines the global score is when all groups have the same error: Any agglomeration function satisfying A2 & A3 will then return the same score. Conversely, the choice of agglomeration can also not be simulated by sensible groupings, as the maximum and the expectation generally behave very differently. These observations suggest that the proposed organisation into grouping and agglomeration is not further reducible.

Intuition (Section 3.4): By suggesting axioms that functions for agglomerating group calibration errors should satisfy, we characterise them as the class of law-invariant coherent risk measures. These can be seen as interpolating between the average and the maximum, which seem to be the only previously suggested agglomeration functions.

13 This definition of CVaR_α shows nicely that it measures the expectation of the largest α% of errors but requires that C has a continuous distribution which is not satisfied in practice, with finite data and groupings. Such a setting requires an alternative definition as $\mathrm{CVaR}_\alpha(C) = \frac{1}{1-\alpha}\int_\alpha^1 Q_C(t)\mathrm{d}t$. A subclass of law-invariant CRMs that subsumes CVaR and is intuitive enough to have been independently 'discovered' by multiple disciplines are the Supremum Risk Measures (Acerbi, 2002; Fröhlich and Williamson, 2024b).

3.5 CALIBRATION-BASED NOTIONS OF FAIRNESS

Having laid out a general framework for designing calibration scores which assess the overall *quality* of probabilistic predictions, we now turn to questions of fairness. Formal scholarship on algorithmic fairness has largely focused on two types of approaches: individual and group fairness. The demand that similar individuals be treated similarly has been proposed as a criterion for **individual fairness** (Dwork et al., 2012). While similar treatment seems to be interpretable as similar predictions, the first 'similar' does all the work here while remaining entirely opaque. An analogous observation has also led some legal scholars to argue that the classic formulation of such a principle in the Western tradition, Aristotle's assertion that like cases should be treated alike, is empty (Westen, 1982). Another problem with such a notion of individual fairness is that it does not take into account the available data in a principled way – it does not use it at all unless the similarity metric is constructed from it. **Group fairness** usually describes the comparison of summary statistics between selected groups, e.g. based on race or gender. One problem with this is that group-level statistics have little to say about the fair treatment of individuals and that they can overlook discrimination across division lines that are unaccounted for. In particular, it reduces individuals to their membership in one particular group that the individual is thought to belong to. One such group-level statistic, proposed in (Kleinberg, Mullainathan and Raghavan, 2017), is calibration, usually evaluated using pre-determined prediction space bins. More recently, scholars have started to bridge the gap from group to individual fairness by considering multitudes of potentially overlapping groups simultaneously, under the name of subgroup fairness (Kearns et al., 2018) or **multi-calibration** (Hébert-Johnson et al., 2018). An open question is the formation of such groups, with the suggestions including all easily computable groups (Hébert-Johnson et al., 2018) or groupings based on similarity of learned representations (Burhanpurkar et al., 2021). Any such grouping is usually subdivided through some prediction-based binning; this is slightly generalised in (Gopalan et al., 2022).

In this section, we connect the richness of calibration laid out in previous sections to the debate on algorithmic fairness. We draw on our characterisation of agglomeration functions for scores of prediction *quality* in Section 3.4 to propose and axiomatise agglomeration functions for global *fairness* scores in Section 3.5.1. In Section 3.5.2, we demonstrate the applicability of global calibration scores based on partitions and localised groups to notions of group and individual fairness, respectively.

3.5.1 *Global fairness scores*

Previous work on calibration for fairness has tended to not discuss global fairness scores, which would allow a direct comparison between algorithms as well as provide an objective to optimise. The original papers (Chouldechova, 2017; Kleinberg, Mullainathan and Raghavan, 2017) only considered bin-wise errors without agglomerating them, while the algorithm proposed in (Hébert-Johnson et al., 2018) runs until the absolute calibration error on every group-bin-combination is below a fixed value, thus implicitly enforcing a constraint on the maximum of all absolute errors. Global fairness scores would need to take into account the signed errors, as we often care much more about whether p over- or under-predicts: When $c(G_1) = \mu$ and $c(G_2) = -\mu$ for some $\mu > 0$, the absolute errors of groups G_1 and G_2 are equal but this is not necessarily fair: When under-prediction of risk is positive from the individuals' perspective, over-prediction is often negative and vice versa. While not every application may have this aspect, it should be possible to distinguish between under- and over-prediction. Furthermore, while adding some constant to all errors changes the quality of the predictor, it does not affect the inequality between the groups.

Caring not about the average calibration error of the groups but about comparisons between them requires measures of deviation rather than risk.[14] However, as we show in Proposition 3.11, there is a close connection between the two. Furthermore, as explained above, we now take 'raw' signed calibration errors as inputs rather than their absolute values. In addition to A3-A5, we suggest that agglomeration functions D for global fairness scores should satisfy the following axiom for $C \in \mathcal{L}^2(J)$:

A6 Normalisation $D(C) \geqslant 0$ and $D(C) = 0 \Leftrightarrow C$ is constant.

Normalisation ensures that the score is positive if and only if there is some inequality in terms of calibration errors. A3-A5 are again desirable for the reasons discussed in Section 3.4. We now introduce a new class of functions.

Definition 3.10 (Fairness Deviation Measures)**.**
A fairness deviation measure (FDM) *is an agglomeration function satisfying A3-A6.*

We can gain some insights into this class by exploiting a handy connection to the law-invariant coherent risk measures discussed in the previous section. This connection is characterised by the following result which is basically a corollary

14 A strand of research that is in some ways similar is that of inequality measures, which have also been proposed for evaluating inequality of losses (Speicher et al., 2018; Williamson and Menon, 2019). In our setting where calibration errors can be positive or negative and have a fixed scale, they seem less suitable.

of the Quadrangle Theorem in (Rockafellar and Uryasev, 2013).[15] It allows the construction of FDMs from CRMs and vice versa.

Proposition 3.11 (Relation between CRMs and FDMs)**.**

a) Agglomeration functions D *satisfying A3-A6 stand in a one-to-one relationship with agglomeration functions* R *satisfying A2-A5 and Aversity (the property that* $\mathsf{R}(C) > \mathbb{E}[C]$ *if* $C \in \mathcal{L}^2(J)$ *is not constant) via*

$$\mathsf{R}(C) = \mathbb{E}[C] + \mathsf{D}(C). \tag{3.12}$$

b) A law-invariant CRM induces an FDM via (3.12) iff it is averse.

c) An FDM induces a law-invariant CRM via (3.12) iff it satisfies $\mathsf{D}(C) \leqslant \sup C - \mathbb{E}[C]$ $\forall C \in \mathcal{L}^2(J)$.

Example 3.12 (Fairness Deviation Measures)**.**

1. *Superquantile deviation:* $\mathsf{D}(C) = \mathsf{CVaR}_\alpha(C - \mathbb{E}[C])$ *for* $\alpha > 0$
2. *Standard deviation:* $\mathsf{D}(C) = \sigma(C)$
3. *Range deviation:* $\mathsf{D}(C) = \sup C - \inf C$

The superquantile deviation is induced by CVaR_α via (3.12). Standard deviation and range deviation do not induce CRMs as they do not satisfy $\mathsf{D}(C) \leqslant \sup C - \mathbb{E}[C]$ $\forall C \in \mathcal{L}^2(J)$. More examples of risk-deviation-pairs can be found in (Rockafellar and Uryasev, 2013). Recall, however, that the coherent risk measures in the previous section are assumed to take absolute calibration errors as inputs while the fairness deviation measures take the 'raw' signed errors.

3.5.2 *Group fairness and individual fairness*

In the remainder of this section, we show that global calibration scores based on different choices of grouping structure (Section 3.3.3) can reflect different notions of fairness. As mentioned above, previous work on calibration and fairness did not provide such population-level scores but largely aimed to compare or minimise absolute group-level errors. The use of global scores based on deviation measures allows a more comprehensive assessment of fairness than previous approaches which compare bin-wise calibration errors between groups of

15 Strictly speaking, the measures must also satisfy that for all $\alpha \in \mathbb{R}$, $\{C \in \mathcal{L}^2(J) : \mathsf{R}(C) \leqslant \alpha\}$ is closed. This is included as a further axiom in (Rockafellar and Uryasev, 2013) but we do not dwell on this since it is automatically satisfied for convex finite measures on finite J, which is always the case in practice.

interest. In case that comparisons between specified, non-overlapping groups are desired, partitions may be the most suitable structure for evaluating calibration as they can provide strict boundaries along such group lines. When based on some form of partition, calibration thus provides a notion **(sub-)group fairness**; this also appears to be in line with much of the previous literature on calibration-based fairness.

In Sections 3.2 and 3.3.1, we pointed to the availability of alternatives to prediction-based grouping. While prediction-based bins for calibration have been suggested as a way to assess whether probabilities 'mean the same thing' (Kleinberg, Mullainathan and Raghavan, 2017), they are at odds with a finer partition based on input features due to constraints on the group sizes (Section 3.3.2). Finer input-based groups, however, allow to avoid more forms of stereotyping and should thus be considered if not as a replacement, then as a complementary basis for evaluating sub-group fairness. A downside of prespecified groups forming a partition is that each individual is only considered as a member of one particular (possibly intersectional) group, which is, after all, a form of pigeonholing. In what follows, we argue that calibration also allows for notions of individual fairness.

A problem with previous (Lipschitz-based) notions of **individual fairness** is that they do not require 'respecting people's claims in proportion to the strength of those claims' (Vredenburgh, 2022). Calibration can offer a principled approach to this by comparing probabilistic predictions with the available data. If we assumed the existence of and access to a true probability distribution $P(Y|X)$ (both of which are problematic), this would provide a measure for the strength of people's claims. Equality of the true calibration error $p(x) - \mathbb{E}_P[Y|X = x]$ across the population would ensure that these claims are respected. However, the theoretical notion of the true calibration error does not provide a notion of individual fairness in practice. We now argue that in practice, calibration on localised groups can provide such a notion.

As shown in Section 3.2, a 'natural' way to construct groups for calibration is as extensions *around* an individual x, in order to make up for limited data *at* x. Furthermore, as established in Propositions 3.5 and 3.6, we can get consistent estimators for the individual calibration error at given points in input space through suitable choices of grouping. Calibration errors of such cohesive groups (or kernels) that can be localised to a particular point in input space and take into account the surrounding datapoints, hence, provide a practical and quantifiable notion of respecting people's claims that is also theoretically appealing. Now recall that the fairness calibration measures introduced above assess the deviation between group-wise calibration errors. This also means that that pres-

ence in multiple or even all groups (Proposition 3.4) does not necessarily imply greater influence on global fairness scores, quite on the contrary: Membership in all groups implies that a datapoint has no influence at all on the global score for a deviation measure like the standard deviation. In sum, global calibration scores for suitable localised groups (e.g. based on k-NN or kernels) can allow to assess whether individuals receive risk scores in proportion to the data available for them, which, we posit, is a notion of individual fairness.

Intuition (Section 3.5): Previous literature has discussed different notions of fairness, notably group and individual fairness. Although calibration is seen as a promising approach to (sub-)group fairness, there is no established metric for it. Potential calibration-based global fairness scores need to take into account the *signed* errors and focus more on their *distribution*. While this implies that they are quite different to the scores discussed in the previous section, they are closely connected. Depending on the grouping structure, global calibration-based fairness scores using deviation measures can provide notions of individual or subgroup fairness.

3.6 CONCLUSION AND FUTURE WORK

By tracing its roots and investigating implicit choices, we have provided a general framework capturing the richness of calibration from the ground up. We paid particular attention to groupings based on the input space and motivated groups that can be localised there. We showed that different ways of evaluating calibration correspond to different notions of fairness, demonstrating the close relationship between calibration and fairness. In the same way as assertions about fairness always need to specify the 'with regard to what?', calibration depends on choices of grouping. For a more solid understanding of this relationship, further formal and conceptual investigations will be necessary. While we made some first steps to analyse the trade-offs inherent in choices of grouping, we see much potential for more comprehensive results. Likewise, while we show a connection between the agglomeration functions suitable for assessing prediction quality on the one hand and fairness on the other, further research on this is required to characterise the prospects of their joint optimisation. To mitigate the dependence on the chosen grouping, it may be beneficial to use multiple different groupings, such as a sequence of increasingly fine partitions; frameworks for utilising such partially redundant information also seem a worthwhile endeavour. In general, we envision our work as a fertile ground for further research on calibration and its role in fairness in particular. We conclude by again stressing

that calibration can only be one element in the just deployment of algorithms in society, given that 'accurate modelling within unjust institutions can replicate and further injustice' (Vredenburgh, 2022).

3.7 APPENDIX

3.7.1 *Established Calibration Scores*

Recall our setting of datapoints $\{(x_i, y_i)\}_{i\in\mathcal{I}}$ with $\mathcal{I} = \{1, ..., N\}$. In (3.4), we defined the calibration error of a group $I \subset \mathcal{I}$ as

$$c(I) := \frac{1}{|I|}\sum_{i\in I} p(x_i) - y_i. \tag{3.13}$$

We now demonstrate how different notions of calibration can be seen as some form of grouping combined with some form of agglomeration of errors. For this, it is useful to think in terms of a parametrised set of groups $\{I_j\}_{j\in J}$ for some index set J which can take different forms. As a first example, we can write the ECE described in (3.5) for bins $B_1, ..., B_K$ of $[0, 1]$ as

$$\mathrm{ECE}(p) = \frac{1}{N}\sum_{j=1}^{N} |c(I_j)|, \quad I_j := \{i \in \mathcal{I} : p(x_i) \in B_{k(j)}\}, \tag{3.14}$$

with $B_{k(j)}$ denoting the bin in which $p(x_j)$ falls. Note that the weighting by the size of the bins is now done implicitly by adding one term for each datapoint, rather than one term per bin. We could avoid this implicit weighting by choosing a different grouping which nevertheless includes the same groups, just in different multiplicities. A score that weights each bin equally, we may call it average calibration error (ACE), is given by

$$\mathrm{ACE}(p) = \frac{1}{K}\sum_{k=1}^{K} |c(I_k)|, \quad I_k := \{i \in \mathcal{I} : p(x_i) \in B_k\}. \tag{3.15}$$

Using the same groups I_k, we can also define the sometimes invoked maximum calibration error (MCE; Naeini, Cooper and Hauskrecht, 2015) in this framework as

$$\mathrm{MCE}(p) = \max_k |c(I_k)|. \tag{3.16}$$

Interestingly, the well-known Brier score is also equivalent to a function over calibration errors for a suitable grouping. As shown in (Sanders, 1963) for a finite prediction space, the Brier score can be decomposed into two terms corresponding to notions of calibration and refinement/sharpness, respectively:

$$\frac{1}{N}\sum_{i=1}^{N}[p(x_i) - y_i]^2 = \sum_{v\in\mathcal{Y}_D} \hat{P}(v)[\tilde{p}(v) - v]^2 + \sum_{v\in\mathcal{Y}_D} \hat{P}(v)\tilde{p}(v)(1 - \tilde{p}(v)) \tag{3.17}$$

where $y_D := \{p(x_i) : i \in \mathcal{I}\}$ is the set of predictions on the observed datapoints, $\hat{P}(X, Y, p(X))$ denotes the empirical distribution on $\mathcal{X} \times \{0, 1\} \times [0, 1]$, and $\tilde{p}(v) := \hat{P}(Y = 1 | p(X) = v)$ denotes the observed ratio of positive labels for datapoints where p predicts v. This can be seen as grouping by level sets of p. Instead grouping the datapoints by inputs allows the decomposition

$$\frac{1}{N}\sum_{i=1}^{N}(p(x_i) - y_i)^2 = \sum_{x \in \mathcal{X}_D} P(x)\left[(p(x)-1)^2 p^\star(x) + (p(x)-0)^2(1-p^\star(x))\right] \tag{3.18}$$

$$= \sum_{x \in \mathcal{X}_D} P(x)\Big[p(x)^2 p^\star(x) - 2p(x)p^\star(x) + p^\star(x) + p(x)^2 - p(x)^2 p^\star(x)\Big] \tag{3.19}$$

$$= \sum_{x \in \mathcal{X}_D} P(x)\Big[p(x)^2 - 2p(x)p^\star(x) + p^\star(x)^2 - p^\star(x)^2 + p^\star(x)\Big] \tag{3.20}$$

$$= \sum_{x \in \mathcal{X}_D} P(x)[p(x) - p^\star(x)]^2 + \sum_{x \in \mathcal{X}_D} P(x)p^\star(x)[1 - p^\star(x)] \tag{3.21}$$

with $P(X, Y)$ denoting the empirical distribution on $\mathcal{X} \times \{0, 1\}$ and $p^\star(x) := \mathbb{E}_P(Y|X = x)$. Note the second term of (3.21) is independent of the predictor p, implying that minimising the Brier score is equivalent to minimising the first term.[16] This first term can be seen as a function of calibration errors:

$$\frac{1}{N}\sum_{i=1}^{N} c(I_i)^2, \quad I_i := \{i \in \mathcal{I} : x_i = x\}. \tag{3.22}$$

The average-of-squares does not satisfy A2 and A3 (Section 3.4), which may have contributed to the observation that 'still there is no clear consensus on how to interpret the Brier score' (Cabitza, Campagner and Famiglini, 2022, p. 85).

As suggested in Section 3.3.3, we can capture notions of calibration that invoke kernels through the generalised calibration error defined for distributions q on $\mathcal{X}$ as

$$C(q) := \mathbb{E}_q[p(X) - p^\star(X)]. \tag{3.23}$$

Kernels have been proposed on the prediction space (Kumar, Sarawagi and Jain, 2018) and on a learned representation space intersected with bins in the

16 This tells us that the Brier score cannot reduce to zero even for the Bayes optimal predictor $p = p$, unless $\forall x \in \mathcal{X} : p^\star(x) \in \{0, 1\}$. This makes sense since the terms $(p(x_i) - y_i)^2$ can only be all zero if no x is observed with both labels, i.e. $\mathbb{E}_P(Y|X = x)$ only takes values in $\{0, 1\}$. A comparison with (3.38) reveals that the second term is the calibration error variance of a point drawn randomly from the empirical distribution!

prediction space (Luo et al., 2022). The maximum local calibration error (MLCE) proposed in (Luo et al., 2022), which we briefly discussed in Section 3.3.3, can in our framework be defined as

$$\mathrm{MLCE} = \max_{x \in \mathcal{X}_{\mathcal{D}}} \mathrm{C}(q_x) \quad \text{for} \quad q_x(x') \propto k_\gamma(x, x') \cdot \mathbf{1}_{p(x') \in B_{k(x)}}(x') \tag{3.24}$$

where k_γ is some kernel on a learned representation space with bandwidth γ and $B_{k(x)}$ denoting the bin (within a pre-defined partition of $[0, 1]$) into which $p(x)$ falls. Their local calibration error (LCE) is simply the generalised group calibration error $\mathrm{C}(q_x)$.

3.7.2 *Some Basic Topology*

Topology is the study of geometric properties such as continuity and connectedness in most general terms. The basic mathematical object is a topological space which is a set equipped with a particular set structure.

Definition 3.13 (Topology, Topological space, Open set, Neighbourhood)**.**
A topology *on a set $\mathcal{X}$ is a collection of subsets $\mathcal{T} \subset 2^{\mathcal{X}}$ satisfying*

- $\mathcal{X}, \emptyset \in \mathcal{T}$.
- *Any union of elements in $\mathcal{T}$ is also in $\mathcal{T}$.*
- *Any finite intersection of elements in $\mathcal{T}$ is also in $\mathcal{T}$.*

The tuple $(\mathcal{X}, \mathcal{T})$ (sometimes just $\mathcal{X}$) is then called a topological space*. The elements of $\mathcal{T}$ (which are sets in $\mathcal{X}$) are called* open sets*. A set that includes an open set containing $x \in \mathcal{X}$ is called a* neighbourhood *of* x.

Example 3.14 (Trivial topology)**.**
For any set $\mathcal{X}$, the trivial topology *is given by $\{\mathcal{X}, \emptyset\}$.*

Example 3.15 (Induced topology, Standard topology on $\mathbb{R}^d$)**.**
For any space $\mathcal{X}$ equipped with a metric d, the topology induced *by d is the topology generated by all balls $B_\epsilon(x)$ with $x \in \mathcal{X}$, $\epsilon > 0$. That is, it contains all sets generated by unions and finite intersections of such balls as well as $\emptyset$ and $\mathcal{X}$. In the case of $\mathcal{X} = \mathbb{R}^d$, all L^r-based metrics induce the same topology which is called the standard topology. For $\mathcal{X} = \mathbb{R}$ in particular, open sets in this topology are exactly the open intervals plus their unions and $\emptyset$.*

Definition 3.16 (Hausdorff space)**.**
A topological space $(\mathcal{X}, \mathcal{T})$ is called a Hausdorff space *if any two points have non-intersecting neighbourhoods.*

Any metric space is Hausdorff under the induced topology while any space with at least two points is not Hausdorff when equipped with the trivial topology.

Definition 3.17 (Topological continuity)**.**
A function between two topological spaces is continuous *if the preimage of each open set is an open set.*

This topological notion of continuity coincides with the better-known ϵ-δ-notion of continuity when the topologies are induced by metrics.
An important notion in general topology and also in the present work is that of connectedness.

Definition 3.18 (Connectedness)**.**
A set in a topological space is connected *if it is not the union of two disjoint nonempty open sets.*

A perhaps more intuitive notion of cohesion is path-connectedness.

Definition 3.19 (Path-connectedness)**.**
A path *from a point x to a point x' in a topological space $\mathcal{X}$ is a continuous function $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x$ and $\gamma(1) = x'$. A set in a topological space is* path-connected *if for any two points in it there is a path between them.*

Path-connectedness is weaker than connectedness: Every path-connected set is connected but not every connected set is path-connected. We do not provide a proof here as this is a standard result. Intuitively, path-connectedness requires a path while connectedness only requires the absence of an open partition. Sometimes you can neither construct two open, separating sets nor a path, in which case we have connectedness but not path-connectedness.

3.7.3 *Proofs*

Proof of Proposition 3.1 (Cohesion in $[0,1]$ and in $\mathcal{X}$)

a) It is a standard result in general topology that the image of a connected set under a continuous map is also connected. A counterexample to the inverse direction is $\mathcal{X} = [-1,1]$ and $p : x \mapsto x^2$ with the standard topologies on $[-1,1]$ and $[0,1]$: the preimage of the connected set $[\frac{1}{4},1]$ is $[-1,-\frac{1}{2}] \cup [\frac{1}{2},1]$ which is not connected.

b) Take the trivial topology $\mathcal{T} = \{\emptyset, \mathcal{X}\}$ on $\mathcal{X}$, take two distinct values $v_1, v_2 \in [0,1]$ that p realises on $\mathcal{X}$. Then take any open interval I around v_1 which

does not contain v_2 (which is possible since $v_1 \neq v_2$ and $[0,1]$ with the standard topology is Hausdorff). Its preimage $p^{-1}[I]$ is neither $\emptyset$ nor $\mathcal{X}$, so not open in the trivial topology. Hence, p is not continuous.

c) Assume p has two (distinct) local maxima at x_a and x_b, with $p(x_a) \leqslant p(x_b)$ (the proof for two minima is analogous). We operationalise the notion of a local maximum x_a in a Hausdorff space as the property that any connected, non-singleton neighbourhood of x_a contains a point x such that $p(x) < p(x_a)$. Now consider the interval $I := [p(x_a), p(x_b)] \subset [0,1]$ under p. If its preimage were connected, there would be a point $x \in p^{-1}[I]$ s.t. $p(x) < p(x_a)$ since it contains the local maximum x_a (which is distinct from $x_b \in p^{-1}[I]$ by assumption). Being the preimage of I, this would mean that $x \in [x_a, x_b]$ and thus $p(x) \geqslant p(x_a)$ which is a contradiction. Therefore, $p^{-1}[I]$ cannot be connected. □

Proof of Proposition 3.2 (Advantage of smaller groups: More information)

a) For $j \in \{1,2\}$, let $z_j := \frac{1}{|I_j|}\sum_{i\in I_j} y_i$ be the average label in I_j. Note that this implies $\sum_{i\in I_j} z_j - y_i = 0$ $(*)$.
If $z_1 \neq z_2$, then the constant predictor defined by

$$p(x) := \frac{1}{|I_1|+|I_2|}\sum_{i\in I_1\cup I_2} y_i \tag{3.25}$$

is perfectly calibrated on the union but not on the individual sets. So from now on now assume $z_1 = z_2$.
Then by assumption, $0 < z_1 < 1$ (as otherwise, all labels would be equal). So there exists $\epsilon > 0$ s.t. $0 < z_1 - \epsilon < z_1 + \epsilon < 1$.
Define $\mathcal{X}_j := \{x_i : i \in I_j\} \subset \mathcal{X}$ for $j \in \{1,2\}$ and choose a p such that

$$p(x) := \begin{cases} z_1 + \frac{|I_2|}{|I_1|+|I_2|}\epsilon & \text{if } x \in \mathcal{X}_1 \\ z_1 - \frac{|I_1|}{|I_1|+|I_2|}\epsilon & \text{if } x \in \mathcal{X}_2. \end{cases} \tag{3.26}$$

Then

$$\frac{1}{|I_1\cup I_2|}\left|\sum_{i\in I_1\cup I_2} p(x_i) - y_i\right| \tag{3.27}$$

$$= \frac{1}{|I_1|+|I_2|}\left|\sum_{i\in I_1}\left[z_1 + \frac{|I_2|}{|I_1|+|I_2|}\epsilon - y_i\right] + \sum_{i\in I_2}\left[z_1 - \frac{|I_1|}{|I_1|+|I_2|}\epsilon - y_i\right]\right| \tag{3.28}$$

$$= \frac{1}{|I_1|+|I_2|}\left||I_1|\cdot\frac{|I_2|}{|I_1|+|I_2|}\epsilon - |I_2|\cdot\frac{|I_1|}{|I_1|+|I_2|}\epsilon\right| = 0, \tag{3.29}$$

using $(*)$ whereas for $j \in \{1, 2\}$,

$$\frac{1}{|I_j|}\left|\sum_{i\in I_j} p(x_i) - y_i\right| = \frac{1}{|I_j|}\left|\pm\frac{|I_1|\cdot|I_2|}{|I_1|+|I_2|}\epsilon\right| > 0. \tag{3.30}$$

b) Assume $|c(I_1)| = \frac{1}{|I_1|}\left|\sum_{i\in I_1} p(x_i) - y_i\right| \leqslant \alpha$ and $|c(I_2)| = \frac{1}{|I_2|}\left|\sum_{i\in I_2} p(x_i) - y_i\right| \leqslant \beta$. Then,

$$\begin{aligned}
|c(I_1 \cup I_2)| &= \frac{1}{|I_1 \cup I_2|}\left|\sum_{i\in I_1\cup I_2} p(x_i) - y_i\right| && (3.31)\\
&= \frac{1}{|I_1|+|I_2|}\left|\sum_{i\in I_1}[p(x_i) - y_i] + \sum_{i\in I_2}[p(x_i) - y_i]\right| && (3.32)\\
&\leqslant \frac{1}{|I_1|+|I_2|}\left(\left|\sum_{i\in I_1}[p(x_i) - y_i]\right| + \left|\sum_{i\in I_2}[p(x_i) - y_i]\right|\right) && (3.33)\\
&\leqslant \frac{\alpha\cdot|I_1| + \beta\cdot|I_2|}{|I_1|+|I_2|}. && (3.34)
\end{aligned}$$

□

Proof of Proposition 3.3 (Advantage of larger groups: Variance)

For a specified point $x \in \mathcal{X}$, we get

$$\mathbb{E}_{P(Y|X=x)}[p(x) - Y] = [1 - p(x)]p^\star(x) + p(x)[1 - p^\star(x)] = p(x) - p^\star(x) \tag{3.35}$$

and, with the same steps as in derivation (3.18)-(3.21),

$$\mathbb{E}_{P(Y|X=x)}[(p(x) - Y)^2] = [p(x) - p^\star(x)] + p^\star(x)[1 - p^\star(x)]. \tag{3.36}$$

This gives

$$\begin{aligned}
\mathbb{V}_P[p(X_1) - Y_1] &= \mathbb{E}_P[(p(X_1) - Y_1)^2] - \mathbb{E}_P[p(X_1) - Y_1]^2 && (3.37)\\
&= \mathbb{E}_P[p^\star(X_1)(1 - p^\star(X_1))]. && (3.38)
\end{aligned}$$

Now since the $(X_1, Y_1), ..., (X_K, Y_K)$ are i.i.d., we can compute

$$\mathbb{V}_P\left[\frac{1}{K}\sum_{i=1}^{K} p(X_i) - Y_i\right] = \frac{1}{K}\mathbb{V}_P[p(X_1) - Y_1] = \frac{1}{K}\mathbb{E}_P[p^\star(X_1)(1 - p^\star(X_1))]. \tag{3.39}$$

□

Proof of Proposition 3.4 (Group overlap for $\mathcal{X}$-based k-NN)

First assume the case $N = 2 \cdot d \cdot (k-1) + 1$.
We now construct a dataset where there are $(k-1)$ points between $\mathbf{o}$ and $\pm\mathbf{e_s}$ for each dimension $1 \leqslant s \leqslant d$, with $\mathbf{e_s}$ denoting the standard unit vector of $\mathbb{R}^d$ along dimension s, except that in dimension d, one point is at $-100 \cdot \mathbf{e_d}$.
For $1 \leqslant s \leqslant d$ (the dimensions of $\mathbb{R}^d$) and $(k-1) \cdot (s-1) < i \leqslant (k-1) \cdot s$ (i.e. $k-1$ points for each dimension s), let x_i be given by $\mu \cdot \mathbf{e_s}$ for some (equal or varying) $0 < \mu < 1$.
For (the following $(k-1) \cdot d - 1$ indices) $(k-1) \cdot d < i < 2 \cdot (k-1) \cdot d$, let $x_i := -x_{i-(k-1)\cdot d}$ and let x_{N-1} be given by $-100 \cdot \mathbf{e}_d$. Lastly, let $x_N := \mathbf{o}$.
Then the point x_N is among the k nearest points for every point (as $\|x_N - x_i\|_p = \|x_i\|_p \ \forall i \in \mathcal{I}$) whereas x_{N-1} is among the k nearest points only of itself.
For the case $N < 2 \cdot d \cdot (k-1) + 1$, we assign less points between $\mathbf{o}$ and $\pm\mathbf{e_s}$, such that x_N is still part of all groups while x_{N-1} is still furthest from all points.
For the case $N > 2 \cdot d \cdot (k-1) + 1$, we assign the additional points x_i for $2 \cdot (k-1) \cdot d \leqslant i < N-1$ between $2 \cdot \mathbf{e_1}$ and $3 \cdot \mathbf{e_1}$ which changes neither the membership of x_N among the k-nearest points of all x_i for $1 \leqslant i < 2 \cdot (k-1) \cdot d$, nor the fact that x_{N-1} is furthest from all points. □

Proof of Proposition 3.5 (Strong consistency of $\mathcal{X}$-based k-NN)

We can reduce our problem to the consistency of k-NN estimators; in the following we adapt notation used in (Devroye et al., 1994):
For fixed p, define $Z := p(X) - Y$. Then Z is bounded and we get a joint distribution (X, Z), a true calibration error $m(x) = \mathbb{E}[Z|X = x] = p(X) - p^*(x)$ and an estimator $m_N(x) := c(I_x^k) = \frac{1}{k} \sum_{i \in I_x^k} z_i$ with $z_i = p(x_i) - y_i$.
Then by Theorem 1 from (Devroye et al., 1994), we get $\mathbb{E}\left[|m_N(x) - m(x)|\right] \to 0$ with probability 1 as $N \to \infty$, which is our desired result.[17] □

Proof of Proposition 3.6 (Strong consistency of $\mathcal{X}$-based kernels)

We can reduce this problem to the consistency of kernel estimators; in the following, we adapt the notation used in (Greblicki, Krzyżak and Pawlak, 1984), analogously to the proof of Proposition 3.5:
For fixed p, define $Z := p(X) - Y$. Then Z is bounded and we get a joint distribution (X, Z), a true calibration error $m(x) = \mathbb{E}[Z|X = x] = p(x) - p^*(x)$ and an

17 Note that we are glossing over the details of tie-breaking, as the expectation is also an expectation over a supporting variable Z.

estimator $m_N(x) := C(q_x^\gamma) = \sum_{i=1}^N q_x^\gamma(x_i) \cdot z_i$ with $z_i = p(x_i) - y_i$.
Additional constraints on k are that

$$c_1 \cdot H(\|x\|_s) \leqslant k(x) \leqslant c_2 \cdot H(\|x\|) \quad \text{and} \quad c_3 \cdot \mathbf{1}_{\|x\| \leqslant r}(x) \leqslant k(x) \tag{3.40}$$

for c_1, c_2, c_3, r positive constants, and H a bounded decreasing Borel function in $[0, \infty)$ such that $t^d H(t) \to 0$ as $t \to \infty$.
Then by Corollary 1 from (Greblicki, Krzyżak and Pawlak, 1984), we get $\mathbb{E}\left[|m_N(x) - m(x)|\right] \to 0$ with probability 1 as $N \to \infty$, which is our desired result. □

Proof of Proposition 3.11 (Relation between CRMs and FDMs)

The Quadrangle Theorem of (Rockafellar and Uryasev, 2013) is stated in terms of two classes of agglomeration functions, for which we require the following further axiom for agglomeration functions R and $C \in \mathcal{L}^2(J)$:

A7 Agreement $C = \alpha \in \mathbb{R}$ constant $\Rightarrow R(C) = \alpha$.

Definition 3.20 (Regular Risk Measures)**.**
A regular risk measure (RRM) *is an agglomeration function satisfying A7, Aversity and Convexity (cf. footnotes 12 and 15).*

Definition 3.21 (Regular Deviation Measures)**.**
A regular deviation measure (RDM) *is an agglomeration function satisfying A6 and Convexity.*

Now Rockafellar and Uryasev (2013) establish that

(A) A convex agglomeration function satisfies A2 if and only if it satisfies A7.

(B) Regular deviation measures D stand in a one-to-one relationship with regular risk measures R via

$$R(C) = \mathbb{E}[C] + D(C). \tag{3.41}$$

(C) R satisfies A3 if and only D does.

(D) R satisfies A1 if and only if D satisfies $D(C) \leqslant \sup C - \mathbb{E}[C]$ $\forall C \in \mathcal{L}^2(J)$.

Furthermore, Artzner et al. (1999) establishes that

(E) An agglomeration satisfying A3 is convex if and only if it satisfies A4.

Lastly, it is easy to see that

(F) For an RRM R and its corresponding RDM D, if one is law-invariant, then so is the other.

Now we can prove Proposition 3.11:

a) An agglomeration function D satisfying A3-A6 satisfies Convexity by (E) and is thus an RDM. An agglomeration function R satisfying A2-A5 and Aversity satisfies A7 by (A) and Convexity by (E) and is thus an RRM. The result follows from (B), (C), and (F).

b) This result follows directly from a) via the definitions of CRMs and FDMs.

c) This result follows from a) and (D).

□

ACKNOWLEDGEMENTS

Big thanks to Rabanus Derr and Christian Fröhlich for helpful discussions and feedback on an earlier version.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645 and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Benedikt Höltgen.

4

A UNIFYING PERSPECTIVE ON PROBABILITIES AS MODEL PREDICTIONS

No one to follow
And nothing to teach
Except that the goal
Falls short of the reach

— Leonard Cohen, *The Goal*

AUTHOR CONTRIBUTIONS

This is a single-author paper.

ABSTRACT

Although probabilistic statements are ubiquitous, foundational disagreements persist about their understanding, as exemplified by debates between Bayesians and frequentists; moreover, it is unclear when and why acting on them actually leads to desirable outcomes. Here, we argue that every probability is the output of a *prediction method*, that is, it depends on both a particular way of constructing abstractions and a way of transforming them into predictions. Through this, we provide a unifying perspective on supposedly different kinds of probabilities and show that even supposedly objective ones are model-dependent. We demonstrate that when a finite calibration criterion is met, one can anticipate the distribution of utilities for a given policy and inform successful decision-making on finite sets of events. Based on the notion of prediction methods, inductive arguments, and the probability calculus, we explain the feasibility of the calibration criterion in many settings. Overall, we develop a coherent perspective on probabilities and their use, connecting key intuitions behind other interpretations along the way.

4.1 INTRODUCTION

> Probabilities are as various as the faces to be seen at will in fretwork or paperhangings
>
> — George Eliot (1871): *Middlemarch*

We make probabilistic statements and use probabilistic reasoning all the time: If the predicted 'probability of rain' is sufficiently high, you bring your umbrella or even stay at home. You decide to undergo surgery if this is thought to significantly 'raise your chances' of recovery. Given that such probabilistic statements permeate both science and our everyday lives, it is quite remarkable that it is still an open question what exactly we mean by them and how they are useful: Do they refer to degrees of belief, to relative frequencies of repeated trials, or to physical properties? The meaning of probability is considered to 'bear at least indirectly, and sometimes directly, upon central scientific, social scientific, and philosophical concerns' (Hájek, 2019). In machine learning, the meaning of probability is increasingly recognised as a 'pressing question' (Burhanpurkar et al., 2021); as put by Cynthia Dwork, 'without an answer to this definitional question, we don't even know what it is that the ideal algorithm should satisfy' (Dwork, 2022).

A common view in both statistics and philosophy is that there are two kinds of probabilities, which one may refer to as aleatory and epistemic, respectively (Hacking, 1975, p. 13–15). Aleatory probabilities are grounded in the world, and potentially objective, for example in gambling or in other observable frequencies. Epistemic probabilities are more speculative, linked to uncertain predictions and credences.[1] In philosophy, these concepts are often linked through the 'Principal Principle', stating roughly that once a chance is learned, it should be adopted as credence. Similarly, it is often assumed 'that the job of statistics is to identify the data-generating process' (Vovk and Shafer, 2025, p. 160), or that in machine learning, 'we would like to match the true data-generating distribution' (Goodfellow, Bengio and Courville, 2016, p. 130).

In contrast to these views, we develop a perspective that unifies these supposedly different kinds of probabilities, arguing that all probabilities are constructed and model-dependent. In this work, we lay out a descriptive account of probability as predictions that are useful when they are finitely calibrated on certain sets. We do not argue for a specific normative position about which predictions should be allowed to be called probabilities; however, our pragmatic

1 Hacking (1975) shows that this distinction is very old and can be found, e.g., in the distinction between *chance* and *probabilité* in the works of Poisson (1837) and Cournot (1843).

perspective can shed new light on common notions of rational belief (abiding by the probability calculus) and decision-making (maximising expected utility).

The paper is structured as follows. In Section 4.2, we introduce the notion of prediction methods and show that they cover not only obvious examples like rain forecasts but also supposedly objective probabilities such as relative frequencies and gambling odds. In Section 4.3, we demonstrate how predictions that are finitely calibrated on relevant sets help us to make actually good decisions. In Section 4.4, we elucidate why calibration is often feasible by drawing connections to the problem of induction and the probability calculus. Lastly, we turn to to the literature on interpretations of probability and argue that our account satisfies general desiderata (Section 4.5) and captures key intuitions behind other interpretations (Section 4.6).

4.2 BEHIND EVERY PROBABILITY IS A PREDICTION METHOD

At the heart of our perspective on probability is the insight that each probability comes from a prediction method. We first introduce the notion of prediction methods with an intuitive case and then walk through two perhaps less intuitive, seemingly aleatory examples.

4.2.1 *Predictors and prediction methods*

All probability assignments involve predictions based on abstractions; the overall process of arriving at such a prediction we call a prediction method. We distinguish the *predictor*—the model that can be seen as a (mathematical) function—from the *prediction method* that is applied to an actual situation. The latter involves the construction of an abstraction (potentially including measurements) before applying the former (Figure 4.1). In the case of rain forecasts, the prediction method consists in first taking measurements (of temperatures, air pressure, etc.) and then feeding them to a computer model (the predictor) that outputs a prediction for the occurrence of rain.

Definition 4.1 (Predictor)**.**
A predictor *is a function* $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *on some set* $\mathcal{X}$ *and algebra*[2] $\mathcal{A}$*.*

2 An *algebra* over a set Ω is a set of subsets of Ω that includes both Ω and the empty set and is closed under complements, finite unions, and finite intersections; this matters for Kolmogorov's axioms (Section 4.4.2).

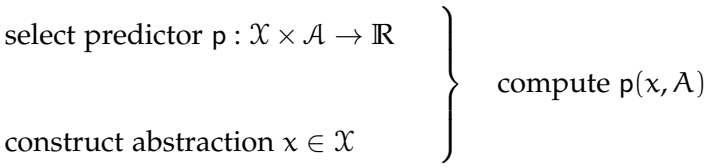

FIGURE 4.1: A prediction method gives a prediction for an event A in some situation by selecting a predictor $p : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with $A \in \mathcal{A}$, constructing an abstraction $x \in \mathcal{X}$ of the situation, and computing $p(x, A)$. For example, p can be a computer model that predicts the event 'rain' based on measurements x.

Definition 4.2 (Prediction method)**.**
A prediction method *for an event $A \in \mathcal{A}$ is an implicit or explicit scheme for selecting a predictor $p : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and constructing an abstraction $x \in \mathcal{X}$ of the given situation.*

The predictor takes two arguments: the abstraction on which to base the prediction, and the event to predict. In many settings, one of the two arguments is effectively ignored. For rain forecasts, the algebra of events $\{\emptyset, \{\mathtt{rain}\}, \{\neg\mathtt{rain}\}, \{\mathtt{rain}, \neg\mathtt{rain}\}\}$ is only implicit. By specifying a prediction p_i for $\{\mathtt{rain}\}$, we intuitively assign predictions $0, (1 - p_i)$, and 1 to the other events, respectively; we will turn to this in Section 4.4.2. Note that the specification of events $A \in \mathcal{A}$ also involves choices of definition or measurement. For example, how much rain counts as 'no rain' or in which area it is recorded is more or less implicit—and depends on value judgements (Douglas, 2000). These choices can, however, be seen as external to the choice of prediction method (although the availability of methods can influence the choice of target), so we do not discuss them further. Importantly, our notion of prediction does *not* require that the predicted events lie in the future, it suffices that the observations/labels are not available to the predictor.

In the rain forecasting example, the abstraction made by the prediction method consists in taking specific measurements of temperature, air pressure, et cetera. More generally, any sort of prediction requires focusing on a subset of all the information that could be taken into account—an abstraction of the situation. Any given situation has an enormous amount of potentially relevant information; typically, we decide what to look at based on experience as well as common sense or expert knowledge. For rain forecasts, temperature, and air pressure are more interesting quantities than the current GDP. Different models for rain prediction (the predictors) can also be based on different ways of measuring temperature—for example, different granularity, location, and timing of the measurements. Different models can also work very differently—they may rely on simple look-up tables or sophisticated simulations. They may even rely on human forecasters who also only use limited information for their forecasts.

While it is difficult to speak of human predictors as stable mathematical objects, they can arguably be approximated as such.

4.2.2 *Example: Symmetry-based predictions*

In many settings, probabilities are intuitively not thought to depend on modelling choices or a particular prediction method; this includes probabilities for gambling devices. Assume you go to a casino where they offer a novel game based on a symmetrical 8-sided and a symmetrical 20-sided 'die' (an octahedron and an icosahedron). Given that it is an official casino, you assume that the dice are indeed symmetrical. How do you make predictions? You can represent any possible outcome that you wish to predict as the set of admissible combinations of faces, which can e.g. be represented as the algebra $\mathcal{A} = 2^{\{1,\dots,8\}\times\{1,\dots,20\}}$. For the prediction, you presumably ignore the name of the croupier, the surface of the table, and so on, and only focus on the symmetry of the dice. Each face of the octahedron corresponds to a prediction of $\frac{1}{8}$ and each face of the icosahedron corresponds to a prediction of $\frac{1}{20}$. How exactly this reasoning is captured by a predictor $p : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is under-determined; in particular, what counts as an argument versus as a part or parameter of the predictor: You may consider a predictor that only takes symmetrical dice, such that the input space $\mathcal{X} = \{x\}$ can be ignored. Alternatively, you may take $\mathcal{X} = \Delta_8 \times \Delta_{20}$ to be the set of all possible combinations of potentially biased 8-sided and 20-sided dice, of which you consider the element $x := (\frac{1}{8}, \dots, \frac{1}{8}, \frac{1}{20}, \dots, \frac{1}{20}) \in \mathcal{X}$, reflecting your assumption of fairness. You may also take a still larger $\mathcal{X}$ that can also capture dice with other numbers of faces. In any case, you proceed by transforming your abstraction of fair dice into a number in $[0, 1]$ through a combinatorial model p. You construct an abstraction and you calculate.

Such symmetry-based prediction methods in gambling situations typically assume that the device is 'fair'. Indeed, gambling devices are produced in such a way that each outcome should occur equally often, which allows us to make roughly accurate predictions about how often events occur on large samples.[3] We know from experience that the process of rolling a fair die is so opaque and chaotic that it is practically impossible for us to predict better than uniformly. If we were more proficient in discerning minuscule variations in die throws and background conditions (as Laplace's demon would be), we might be able to

3 Hacking (1975, p. 4) notes that already '[t]he dice in the cabinets of the Cairo Museum of Antiquity, which the guards kindly let me roll for a long afternoon, appear to be exquisitely well balanced.' It also appears that probabilistic calculations were already known to gamblers before mathematicians started to take an interest in it in the 17th century (Garber and Zabell, 1979).

make finer predictions. Indeed, Edward Thorp and Claude Shannon developed a device to better predict roulette outcomes which they successfully deployed in casinos in the 1960s (Thorp, 1998).After all, the assumption of a fair die or roulette wheel is a particular abstraction of (your beliefs about) the situation.

4.2.3 *Example: Frequency-based predictions*

Predictions that are explicitly based on looking up relative frequencies of similar events use a very simple type of prediction method. Such a method could be used for gambling instead of (or combined with) symmetry-based considerations. There are also more interesting examples, such as medical risks. Doctors typically base risk predictions on past experience, sometimes by explicitly looking up data about similar people. In an example recently discussed in (Dawid, 2017), which we shall return to later, Angelina Jolie got told that she had an 87% risk of breast cancer—with the number presumably coming from statistical data about women with a particular genetic mutation. Hence, the doctors used the following prediction method: They chose a particular abstraction of Angelina, as a woman with this gene mutation, and then used a predictor that is basically a look-up table. Presumably, women without that gene mutation get assigned into different categories (or 'reference classes') for which there is enough data for doctors to believe that the relative frequency in this category is stable over time. Different doctors may use different categories, that is, different abstractions. As in the rain example, we can consider this as a case where the implicit 4-element algebra is ignored. Alternatively, we could see it as an application of a more general predictor that can output predictions for different diseases, based on multiple look-up tables. This would make $\mathcal{A}$ more complex by adding more diseases as fundamental events and means that $\mathcal{X}$ needs to be fine enough that all relevant categories for all diseases can be distinguished.[4]

4.3 FINITE CALIBRATION MAKES PROBABILITIES USEFUL

The previous section argued that probabilities are outputs of prediction methods and thus constructed; this raises the question why we construct them and how they work. Although it is often assumed that probabilistic predictions are useful for decision-making, it has not been demonstrated in general terms how or under which conditions this is the case. In this section, we show how *finite sets* of predictions are useful to us if they satisfy a form of calibration. Although

4 One may also use a weaker set structure to not allow all intersections (Derr and Williamson, 2023).

the technical details of the general perspective are almost trivial, it seems to not have been discussed before, let alone its relevance appreciated. From this general understanding, the idea of expected utility maximisation and the more common understanding of calibration emerge as special cases. For the purpose of this section, it is enough to think of predictions as arbitrary real numbers $p_i \in \mathbb{R}$.[5] As before, a prediction p_i relates to an event A_i and its label $y_i \in \{0, 1\}$ where $y_i = 1$ or $y_i = 0$ denotes that the event does or does not occur, respectively.

4.3.1 *Predicting numbers of events*

What is the difference between a prediction of 0.6 and a prediction of 0.9, given that the predicted event either does or does not occur? An important difference surfaces when considering multiple predictions: In general, of 100 events with prediction 0.6, we intuitively expect roughly 60 to occur, whereas of 100 events with prediction 0.9, we would expect roughly 90 to occur. We can formalise this as a quality criterion for predictions called *calibration*: For a given set of events, the sum of our predictions should coincide with the number of occurring events.

Definition 4.3 (Calibration)**.**
Predictions $p_1, ..., p_d \in \mathbb{R}$ *are said to be* calibrated *for observations* $y_1, ..., y_d \in \{0, 1\}$ *if they satisfy*

$$\sum_{i=1}^{d} p_i = \sum_{i=1}^{d} y_i. \tag{4.1}$$

Often, calibration is understood more narrowly as what Dawid (2017) calls 'probability calibration', namely calibration *on sets of equal prediction*. While this unnecessarily narrow understanding of calibration may be partly due to historical reasons, as discussed in (Höltgen and Williamson, 2023), it is also particularly relevant in many settings (Section 4.3.3). Our definition is similar to the definition of calibration of Dawid (1985), except that there, it is restricted to infinite sub-sequences of infinite sequences of outcomes whose averages are assumed to converge (which essentially presupposes the existence of objective, frequentist probabilities). We argue that predicting how many events of certain sets will occur, i.e. calibration in the general sense, is the purpose for having probabilities in the first place.

The intuition for this is simple: If I knew that my predictions are calibrated on a set of events, then the sum of the predictions tells me how many of the set

5 We will, however, demonstrate the benefits of satisfying Kolmogorov's axioms in later sections.

of events will occur. This is very helpful, for example, in gambling settings: if I can reliably predict how often a repeatable event with given payoff will occur, then I know how much I should bet in each instance to come out positively in the end. An important aspect here is that I do not care which of the events will occur, because the payout is always the same. This is what calibration delivers: It tells you how many events will occur, without telling you which ones.

An implication of this is that calibration is only useful if I care equally about each event. This notion of 'caring equally' is captured in formal models by the condition that all events give me the same *utility*. Intuitively, a person's utility is a numerical representation of how much the person values the (non-)occurrence of an event (which is clearly an idealisation). In our setting of binary events, we will use utility functions $u_i : \{0,1\} \mapsto \mathbb{R}$ where $u_i(0)$ and $u_i(1)$ capture how much I value $y_i = 0$ and $y_i = 1$, respectively. Now, calibration can help us to foresee the (non-normalised) *distribution of utilities* that I will receive: If my predictions are calibrated on sets of equal utility, then I can predict the number of occurring events for each utility by summing up the relevant predictions.

4.3.2 *Predicting cumulative utility*

While we will come back to general utility distributions in Section 4.3.4, we will for now focus on a particularly intuitive property of that distribution, which we call cumulative utility.

Definition 4.4 (Cumulative utility)**.**
My cumulative utility *over a set of outcomes* $\{y_1, ..., y_d\}$ *given utility functions* $u_1, ..., u_d$ *is*

$$\sum_{i=1}^{d} [y_i u_i(1) + (1 - y_i) u_i(0)] \,. \tag{4.2}$$

As y_i denotes whether event 1 or 0 occurs, the cumulative utility is the *sum over all utilities that I actually receive*. This captures how well I will be off overall. We now formally show that with suitably calibrated predictions, it is possible to estimate cumulative utility through a very familiar quantity. The simple idea is that if for each utility value, I correctly predict how many events with this utility will occur, then I will correctly predict my cumulative utility. Note that for our purposes, it would be mathematically equivalent to consider the average utility I get at each time step, i.e. the cumulative utility divided by d.

Proposition 4.5 (Predicting cumulative utility)**.**
Let there be d *predictions* $p_i \in \mathbb{R}$ *for binary outcomes* $y_i \in \{0,1\}$, $i \in \{1, ..., d\}$ *with*

utility functions $u_i : \{0,1\} \to \mathbb{R}$ *and assume that the predictions are calibrated on sets of equal utility* $u_i(0)$ *and on sets of equal utility* $u_i(1)$*, formalised by the assumption that* $\forall u' \in \mathcal{U}$:

$$\sum_{i:u_i(1)=u'} y_i = \sum_{i:u_i(1)=u'} p_i \qquad \textit{and} \qquad \sum_{i:u_i(0)=u'} y_i = \sum_{i:u_i(0)=u'} p_i, \tag{4.3}$$

where $\mathcal{U}$ *denotes the set of all values that the* u_i *can take, i.e.*

$$\mathcal{U} := \bigcup_{1 \leqslant i \leqslant d} \{u_i(0), u_i(1)\} \subset \mathbb{R}.$$

Then I can correctly predict my cumulative utility (LHS) via

$$\sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] = \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)]\,. \tag{4.4}$$

Proof.

$$\begin{aligned}
&\sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] \\
&\quad = \sum_{u' \in \mathcal{U}} \left(\sum_{i:u_i(1)=u'} y_i + \sum_{i:u_i(0)=u'} (1-y_i) \right) \cdot u' && (4.5)\\
&\quad = \sum_{u' \in \mathcal{U}} \left(\sum_{i:u_i(1)=u'} p_i + \sum_{i:u_i(0)=u'} (1-p_i) \right) \cdot u' && (4.6)\\
&\quad = \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)] && (4.7)
\end{aligned}$$

where for (4.5) = (4.6), we use the calibration criterion (4.3). □

Given that the assumption of *exact* calibration on all sets of *equal* utility is very strong, we also show that approximate calibration (Appendix 4.8.1.1) and calibration on sets of approximately equal utility (Appendix 4.8.1.2) suffice for approximately correct predictions of the cumulative utility. Furthermore, a weaker calibration criterion for imprecise predictions makes it possible to incorporate risk aversion (Appendix 4.8.1.3). Note that the RHS of (4.4) is the sum over my expected utilities (my expected cumulative utility), in the conventional probabilistic framework where $\hat{Y}_i$ is the random variable with distribution P_i that takes value 1 rather than 0 with probability p_i:

$$\sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)] = \sum_{i=1}^{d} \mathbb{E}_{P_i}[u_i(\hat{Y}_i)]. \tag{4.8}$$

This means that the policy of expected utility maximisation (EUM) is the policy that actually maximises my cumulative utility if my predictions are calibrated! To capture the idea of maximising utility, we need a notion of decisions between acts, which is not yet part of the setup. For simplicity, we consider d binary decisions, at step i consisting in a choice between (p_i^a, u_i^a) and (p_i^b, u_i^b).

Corollary 4.6 (Comparing policies by expected utility)**.**
For $i \in \{1, ..., d\}$, let there be predictions $p_i^a, p_i^b \in \mathbb{R}$ for binary outcome $y_i \in \{0, 1\}$ and utility functions $u_i^b, u_i^b : \{0, 1\} \to \mathbb{R}$. For $\pi \in \{a, b\}$, policy π is then given by predictions $p_1^\pi, ..., p_d^\pi$ and utility functions $u_1^\pi, ..., u_d^\pi$. Assume that the predictions of both policies are calibrated on sets of equal utility in the sense of (4.3) for their respective utility functions.
Then, for $\hat{Y}_i^\pi$ and P_i^π as in (4.8), the policy with the higher expected cumulative utility $\sum_{i=1}^d \mathbb{E}_{P_i^\pi}[u_i^\pi(\hat{Y}_i^\pi)]$ will actually provide the higher cumulative utility.

While EUM can be seen as a descriptive theory of human decision-making, it is often also assumed as a normative principle (e.g. by Hedden (2013)). We can now give conditions under which this is a good policy in the sense that it leads to desired outcomes: when we are calibrated on sets of equal utility and when we care about maximising cumulative utility. We now discuss a specific case of utility settings that has received much attention in statistics and machine learning, before we widen the scope again and consider cases where we are not interested in cumulative utility.

4.3.3 *Probability calibration*

We now illustrate the above with an example. Let there be d days where for $i \in \{1, ..., d\}$, $p_i \in \{0, 0.1, 0.2, ..., 1\}$ is the daily rain forecast and $y_i \in \{0, 1\}$ denotes whether it actually rains ($y_i = 1$ denoting rain). Now assume that across days, my attitude towards rain does not change but that it depends on whether I brought an umbrella: Let my utilities be given by $u^a(1) = 0$ and $u^a(0) = -1$ if I brought an umbrella and $u^b(1) = -3$ and $u^b(0) = 0$ if I did not bring one. Then my predicted utility when bringing an umbrella on day i is

$$p_i \cdot u^a(1) + (1 - p_i) \cdot u^a(0) = (1 - p_i) \cdot (-1) = p_i - 1 \tag{4.9}$$

whereas my predicted utility when not bringing an umbrella is

$$p_i \cdot u^b(1) + (1 - p_i) \cdot u^b(0) = -3p_i. \tag{4.10}$$

As $p_i - 1 > -3p_i \Leftrightarrow p_i > 0.25$, I maximise predicted utility (per day) if I bring an umbrella on days where $p_i > 0.25$. This policy induces the utility functions

$$u_i = \begin{cases} u^a & \text{if } p_i > 0.25 \\ u^b & \text{if } p_i < 0.25. \end{cases} \tag{4.11}$$

The calibration criterion (4.3) in Proposition 4.5 for this case amounts to

$$\sum_{i:p_i<0.25} y_i = \sum_{i:p_i<0.25} p_i \qquad \text{and} \qquad \sum_{i:p_i>0.25} y_i = \sum_{i:p_i>0.25} p_i. \tag{4.12}$$

If this condition is satisfied, my cumulative utility will coincide with the sum of my daily predicted utilities.

To assess the relative merits of this particular policy, we need to compare it with other policies. (4.12) is one instance of a prediction-dependent threshold policy, where I bring an umbrella whenever the predicted probability of rain is higher than a certain threshold (in this case, 0.25). For Proposition 4.5 to apply to such a policy with any threshold $t \in [0, 1]$, the calibration criterion is

$$\sum_{i:p_i<t} y_i = \sum_{i:p_i<t} p_i \qquad \text{and} \qquad \sum_{i:p_i>t} y_i = \sum_{i:p_i>t} p_i. \tag{4.13}$$

Now note that since the possible utility functions are assumed to be the same each day, for this condition to be satisfied for all $t \in [0, 1]$, it is enough to satisfy

$$\forall v \in \{0, 0.1, 0.2, ..., 1\} : \sum_{i:p_i=v} y_i = \sum_{i:p_i=v} p_i. \tag{4.14}$$

Now this is just the common condition of calibration on sets of equal prediction, which is commonly expected of rain forecasters (Gigerenzer et al., 2005) and which 'even inexperienced forecasters are capable of displaying', except for extreme predictions (Sanders, 1963, p. 191); see also Murphy and Winkler (1977). Under this fairly benign assumption, choosing 0.25 as my threshold maximises not only my *predicted* utility but also my *actual* cumulative utility among all threshold-based policies due to Proposition 4.5! Hence, people can tailor their policies to their personal utilities and, thus, their decisions to the forecasts. This also demonstrates why probabilistic forecasts are useful even in a deterministic world without 'real' probabilities (cf. Section 4.5.1). We would like to highlight that calibration on sets of equal prediction thus derives its importance (and prevalence) from their concurrence with the sets of equal utility when considering threshold-based policies.

This observation is closely linked to the connection between probability calibration and swap-regret that was shown for the first time by Foster and Vohra

(1998). In this work, a randomised forecasting algorithm which minimises the maximum regret under a permutation of predictions (the maximum swap regret) was used to achieve low calibration error; this was used to show that one can always achieve probability-calibrated forecasts, at least in probability. While the importance of the reverse direction has since been appreciated (Noarov and Roth, 2024), the benefit of calibration is commonly only framed in terms of regret; in our view, the connection to swap regret should be seen as a corollary of the more fundamental function of probability *per se*, the accurate prediction of numbers of occurring events. Closest to ours is the perspective of Zhao et al. (2021) which focusses on the predictability of average loss, but is restricted to loss functions which depend only on prediction and outcome, not on the more general utility of the event.

Note that the calibration criterion was fairly benign in our rain example because the utilities are the same every day and we restricted our comparisons to the 10 threshold-based policies (arguably the only sensible policies here). The story would be more complex if we took the utility to also depend e.g. on wind speed (because it affects the efficacy of umbrellas) or on the day of the week. In general, for d binary decisions, there are 2^d possible combinations of decisions, i.e. policies! Accordingly, if we wanted to compare the cumulative utility of all possible choices via Proposition 4.5, this would lead to a very strong calibration criterion—in fact, it would require perfect binary predictions $p_i = y_i$. This should not be surprising, as the best combination of decisions would be to always bring an umbrella if and only if it rains, which we can only ensure if we can discriminate perfectly between rainy and dry days.

It is, therefore, important to emphasise that calibration on sets of equal prediction should not be the only quality criterion for predictions. Consider the rain example above: While the constant base rate predictor also satisfies the calibration criterion (4.14), more refined predictions would allow people to better tailor their decisions to their utilities. Another property of interest is, thus, what is sometimes called sharpness or refinement, relating to the information content of a predictor (DeGroot and Fienberg, 1983). The Brier score, to take a criterion important in rain forecasting, can be decomposed into two terms measuring probability calibration and sharpness, respectively (Sanders, 1963). These two properties are in tension in the sense that it is more difficult to be calibrated on more informative predictions. While the focus of this work is how predictions can be useful in general, rather than their evaluation, we presently also mention some connections to the latter.

4.3.4 *Calibration revisited*

We now briefly consider alternatives to the cumulative utility as the quantity of interest. Recall from Section 4.3.1 that predictions calibrated on sets of equal utility not only allow us to predict the cumulative (or average) utility but the *distribution* of utilities more generally. Let $\mathcal{D} := \{\mu : \mathbb{R} \to \mathbb{N}_{\geqslant 0}\}$ denote the space of utility distributions, i.e. functions indicating how often different utility values occur.[6] Let $\mathcal{U}_\mu := \{u \in \mathbb{R} : \mu(u) > 0\}$ denote the support of a utility distribution $\mu \in \mathcal{D}$, i.e. the set of utility values that do occur in the distribution μ. With this notation, we can describe the cumulative utility of a distribution $\mu \in \mathcal{D}$ as $\sum_{u \in \mathcal{U}_\mu} u \cdot \mu(u)$. Another potentially relevant property of utility distributions is the smallest received utility, $\min \mathcal{U}_\mu$. For the umbrella policies, optimising for this property would mean that we should always bring an umbrella when $p_i > 0$, as we will otherwise incur a utility of -3 at some point—assuming that the calibration condition (4.14) holds. The minimum is quite an extreme property of a distribution as it ignores most information about the distribution (in our case, all events except those with the lowest utility). Optimising other properties of utility distributions will lead to other decision criteria but these questions are not the focus of this paper, interesting as they are.

In concurrent work, Perdomo and Recht (2025) claim that '[t]he utility of calibration comes in terms of communication' (p. 15) and 'emphasize that beyond [the] property of shared interpretation, calibration doesn't mean much' (p. 18). Against this view, we highlight three interrelated perspectives on the importance of calibration. First, the equivalence of swap regret and calibration highlights that the 'best-response action' which maximises EU will fare better than the reverse policy. In the rain case, even just matching the base rate would already allow to always take the action compatible with the more probable event—which is not impressive but better than the opposite. Arguably, this is only one part of a second, broader perspective on the choice of predictor given some policy and utility/loss: Calibrated predictors guarantee better decision outcomes than miscalibrated predictors with the same discriminative power. This perspective is explored further in many works connecting calibration to loss minimisation (Derr, Finocchiaro and Williamson, 2025; Feng and Tang, 2025; Gopalan, Kim and Reingold, 2023; Kleinberg et al., 2023). The third perspective is closer to our basic idea of predicting utility distributions: If we

6 For $u \in \mathbb{R}$, $\mu(u) = n$ then means that the utility u occurs n times in the distribution described by μ. Instead of such distributions, one may also think of multisets. Note that calibration on sets of equal utility generally only allows us to predict *how often*, but not *when* a specified utility will be received.

assume calibration on relevant sets, we can choose among policies based on the utility distributions they will, respectively, lead to. Höltgen and Williamson (2026a) demonstrate how this perspective can be fruitfully applied to causal inference settings. These perspectives are different sides of the same gambling device; combining the latter two, one may suggest choosing policy and predictor together, based on the promised utility distribution and on how realistic it is that the required calibration conditions are met. This highlights a fundamental question: How can we say anything about whether a set of predictions will be calibrated?

4.4 HOW IS CALIBRATION POSSIBLE?

Even if we showed calibration can make predictions useful, this only helps with understanding probability in the case that is also realistic to achieve calibration. For arbitrary sets of predictions, there need not be a reason to assume that such a criterion would be satisfied. This points to the importance of considering the methods that generated the predictions, which we discussed in Section 4.2. Based on the notion of calibration for predictions $p_1, ..., p_d$ as defined above, we can define calibration for predictors and prediction methods. For this, we consider the predictions of d events $A_1, ..., A_d \in \mathcal{A}$, with a prediction method that uses abstractions $x_1, ..., x_d \in \mathcal{X}$.

Definition 4.7 (Calibration of predictors and prediction methods)**.**
A predictor (or prediction method) is said to be calibrated on events $A_1, ..., A_d \in \mathcal{A}$ *for observations* $y_1, ..., y_d$ *if its predictions* $p_1, ..., p_d$ *are calibrated.*

Prediction methods provide a first way to draw connections between individual predictions, which is necessary to even start talking about satisfying criteria on *sets* of events—but how can we hope for calibration on *unseen* sets of events? A first idea of providing calibrated predictors may be to directly optimise for it, in what has been called defensive forecasting (Vovk, Takemura and Shafer, 2005) or forecast hedging (Foster and Hart, 2021). This literature emerged in response to Schervish (1985) showing that calibration in the classical sense cannot be guaranteed by any algorithm. Foster and Vohra (1998) showed that, under very mild assumptions, probability calibration in the sense of a low ECE can be guaranteed (in probability) by a stochastic algorithm that minimises swap regret (for the Brier score). However, such backward-looking algorithms often simply converge to (more or less stable) base rates, without refined predictions that allow well-informed policies. While they can be adapted to smaller patches of the input space (see also the literature on multi-calibration (Hébert-

Johnson et al., 2018)), this is not much better than simply predicting the average per patch, thereby deciding in advance which patches get individual decisions. In this, such approaches are more constrained than, e.g., human weather forecasters who were trained to first sort similar weather situations into 'categories of likelihood of occurrence' and then predict a calibrated forecast per category (Sanders, 1963, p. 200).

This example and those of Section 4.2 may suggest that induction about calibration always reduces to stable relative frequencies of repeated trials; this is also not the case. Consider simulation models in the rain prediction example: If the measurements are fine enough, it may be the case that no input $x_i \in \mathcal{X}$ occurs more than once and no prediction is issued more than once, implying that there are no repeated trials. Further examples are given by logistic regression or more complex Machine Learning models used for probabilistic predictions. In general, calibrated predictors may rely on some structure in the relationship between inputs and labels that does not reduce to stable relative frequencies. Otherwise, simulation-based and ML-based predictions could simply be replaced by 'reference class forecasting' (Flyvbjerg, Glenting and Rønnest, 2004).

Empirically, we do see that prediction methods can often be designed such that they are (approximately) calibrated on sets of interest, which simply means that they neither systematically over- nor systematically under-predict. Besides rain forecasts and the examples in Sections 4.2.2 and 4.2.3, we can point to machine learning (ML) models which often aim for calibration on sets of equal prediction. It has been observed that especially modern, over-parameterised ML models need explicit post-processing, whereas others are automatically calibrated on sets of equal prediction (Guo et al., 2017). One could argue that on a high level, humans also do something like this post-processing: if we are repeatedly over- or under-predicting (i.e. are not calibrated) on sets of interest, we will ideally notice that; since we do not know on which of the individual events our predictions were too low/high, we systematically increase/decrease our predictions in similar situations[7] in the future. But why should future predictions then still be calibrated?

7 Choosing a sensible notion of similarity here also requires experience. Nelson Goodman (1972, p. 18) already suspected 'that rather than similarity providing any guidelines for inductive practice, inductive practice may provide the basis for some canons of similarity'.

4.4.1 *Induction and the feasibility of calibration*

Any prediction method, indeed any prediction about the future, relies on an inductive assumption: that the future will resemble the past in some relevant way. This relevance can be made more precise for our purpose: that a prediction method which has repeatedly proven to be (approximately) calibrated on some sets in the past will be (approximately) calibrated on similar sets in the future. Below, we prove a formal result in support of this particular inductive assumption, similar to the argument for induction made by Williams (1947). In contrast to the cited work, we are dealing not only with integers but with real numbers, which is why we draw on an established concentration inequality.[8] In particular, we make use of the combinatorial bound provided by Hoeffding's inequality for drawing without replacement.

Proposition 4.8 (Calibration on samples from a population)**.**
Take a predictor $p : \mathcal{X} \times \mathcal{A} \to [0, 1]$ *and* N *prediction instances represented by* $(x_i, A_i) \in \mathcal{X} \times \mathcal{A}$*. Now consider drawing a 'sample' of* d *instances from the 'population' of* N *instances. Then the samples* $\{j_1, ..., j_d\} \subset \{1, ..., N\}$ *whose average calibration error differs by more than* ϵ *from the average calibration error of the population, i.e. where*

$$\frac{1}{d}\sum_{i=1}^{d}\left(p_{j_i} - y_{j_i}\right) - \frac{1}{N}\sum_{i=1}^{N}\left(p_i - y_i\right) \geqslant \epsilon, \tag{4.15}$$

(y_i *denoting whether* A_i *occurs and* $p_i := p(x_i, A_i)$*) make up for less than* $\exp\left(-\frac{1}{2}d\epsilon^2\right)$ *of all possible samples of that size.*

Proof.
Our result follows directly from Hoeffding's inequality for drawing without replacement. We simply insert $z_i := p_i - y_i$, $a = -1$ and $b = +1$ in the below statement taken from Proposition 1.2 of Bardenet and Maillard (2015):

Let $Z = (z_1, ..., z_N)$ be a finite population of N points and $Z_1, ..., Z_d$ be a random sample drawn without replacement from Z. Let

$$a := \min_{1 \leqslant i \leqslant N} z_i \quad \text{and} \quad b := \max_{1 \leqslant i \leqslant N} z_i. \tag{4.16}$$

Then for all $\epsilon > 0$,

$$\mu\left[\frac{1}{d}\sum_{i=1}^{d} Z_i - \frac{1}{N}\sum_{i=1}^{N} z_i \geqslant \epsilon\right] \leqslant \exp\left(-\frac{2d\epsilon^2}{(b-a)^2}\right), \tag{4.17}$$

where μ measures the proportion of admissible combinations in drawing d of the N points. □

8 Also note that we do not mean to provide an a priori justification for induction here: our goal is not to solve any (old or new) riddle of induction but to elucidate the role of probability in it.

Hence, the average calibration error on large enough samples will mostly be close to the average calibration error of the whole population. In a move analogous to that of Williams (1947), we can also infer that if I am approximately calibrated on a large enough sample from a population or set of prediction instances, I will in most cases also be calibrated on the whole set and, thus, on similar sets in the future. Let us illustrate the bound with concrete numbers. If I have an average calibration error of 0.2 on the whole population, then I will get a calibration error of less than 0.05 in less than 10% of possible samples of size $d = 200$; for $d = 500$, this ratio goes down to 0.4%. Hence, the vast majority of possible samples will not mislead me into thinking that I will be well-calibrated in the future in such a setting. Note that the proposition only provides an upper bound, so that the actual number of non-representative samples will be lower still. While this result does not prove the possibility of induction, it shows that calibration in the past is an indicator for calibration in the future—on sets that can be thought to be drawn from the same population. Whether this is a sensible model in a given situation depends on whether there is reason to believe that the sample is unbiased.[9]

Another interesting implication of the result concerns the mixing of predictions from multiple, different calibrated prediction methods. If n prediction methods are calibrated on d events each, then the resulting $n \cdot d$ predictions are clearly also calibrated on the $n \cdot d$ events; Proposition 4.8 can also be applied to this larger set of predictions, now inferring from the population to subsets: It shows that most large enough subsets of these $n \cdot d$ predictions will also be approximately calibrated, even though they come from a mix of prediction methods. This is important because it shows that the Proposition is not only relevant for predictions from the same prediction method. In sum, we can give arguments why, in certain cases, we expect to be calibrated on future events—but we can never be sure: 'Nature will always maintain her rights, and prevail in the end over any abstract reasoning whatsoever' (Hume, 1777, p. 5.1.2).

4.4.2 *Extrapolating calibration*

Another way of generating sets of calibrated predictions is through certain other sets of calibrated prediction—by using probabilistic reasoning. Even attentive readers probably missed the interesting fact in Proposition 4.5 that $1 - p_i$ automatically emerged as the prediction for $1 - y_i$, without imposing Kolmogorov's axioms. We now show more generally that predictors need to satisfy these ax-

9 An illuminating analysis of the difference between the means in terms of the bias of the sampling procedure is given by Meng (2018).

ioms (in their second argument) in order to be calibrated on certain sets.[10] Take a predictor $p : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and d prediction instances represented by $(x_i, A_i) \in \mathcal{X} \times \mathcal{A}$ with $p_i := p(x_i, A_i)$ and let $y_i \in \{0, 1\}$ denote whether A_i occurs. Let $y_A, y_B, y_{A\cup B}, y_\Omega \in \{0, 1\}$ denote whether events $A, B, A \cup B, \Omega \in \mathcal{A}$ occur at the last instance d. Let $\mathcal{A}$ be an algebra over some set Ω where Ω is a sure event: It exhausts all possibilities; that is, for its label, it is known that $y_\Omega = 1$.

1. *Non-negativity:* If there is a nontrivial subset $I := \{i : p_i < 0\} \subset \{1, ..., d\}$ where p predicts negative values, then p cannot be calibrated on this subset, regardless of whether the predicted events occur:

$$\sum_{i\in I} p_i < 0 \leqslant \sum_{i\in I} y_i. \tag{4.18}$$

2. *Normalisation:* Let $A_d = \Omega$ and p be calibrated on the set $\{1, ..., d-1\}$. Then p is calibrated on $\{1, ..., d\}$ if and only if $p(x_d, \Omega) = 1$, regardless of x_d.

3. *Additivity:* Let $A, B \in \mathcal{A}$ be disjoint events in the sense that $y_A + y_B \leqslant 1$ (i.e. it cannot be that both labels are equal to 1 at the same instance); this implies $y_A + y_B = y_{A\cup B}$. Now assume p is calibrated on the set $S := \{(x_1, A_1), ..., (x_{d-1}, A_{d-1}), (x_d, A), (x_d, B)\}$, where p is used for two predictions at instance d. Then

$$\begin{aligned}
&p(x_d, A \cup B) + \sum_{i=1}^{d-1} p_i - y_{A\cup B} - \sum_{C\in J} y_i \\
&\quad = p(x_d, A \cup B) + \sum_{i=1}^{d-1} p_i - y_A - y_B - \sum_{i=1}^{d-1} y_i && (4.19)\\
&\quad = p(x_d, A \cup B) - p(x_d, A) - p(x_d, B) && (4.20)\\
&\qquad + p(x_d, A) + p(x_d, B) + \sum_{i=1}^{d-1} p_i - y_A - y_B - \sum_{i=1}^{d-1} y_i \\
&\quad = p(x_d, A \cup B) - p(x_d, A) - p(x_d, B) && (4.21)
\end{aligned}$$

   where the last step uses the assumption of calibration on S. So under that assumption, p is calibrated on $\{1, ..., d\}$ with $A_d = A \cup B$ if and only if $p(x_d, A \cup B) = p(x_d, A) + p(x_d, B)$, regardless of x_d.

The sets that allow if-and-only-if statements are quite specific here; in this sense, it resembles Dutch book arguments, where any single inconsistency can in theory be exploited indefinitely. Here, however, the implications are more practical:

10 Related observations for the case of sets of equal prediction have been made by Van Fraassen (1983).

If someone is perfectly calibrated on forecasting 'rain' but does not obey the probability axioms on one 'no rain' forecast, then for some utility functions (in the setting of Section 4.3.3), the best-response policy is guaranteed to lead to sub-optimal decisions due to miscalibration.

We can also motivate the definition of conditional probabilities by the demand for calibration (somewhat analogous to definitions via relative frequencies). Consider the task of predicting events $A, B \in \mathcal{A}$ at d instances. For ease of presentation, assume that all d inputs coincide, i.e. $x_1 = ..., x_d = x \in \mathcal{X}$. This allows us to drop p's dependence on $x \in \mathcal{X}$ and consider a predictor $p : \mathcal{A} \to [0,1]$ in the following derivation; a more general version is presented in Appendix 4.8.2. Now assume p to be calibrated on $A \cap B$ and on B across the d instances, where $y_i^{A\cap B}$ and y_i^B denote whether $A \cap B$ and B occur at instance $i \in \{1, ..., d\}$, respectively. That is, assume $\sum_{i=1}^d p(A \cap B) = \sum_{i=1}^d y_i^{A\cap B}$ and $p(B) = \frac{1}{d}\sum_{i=1}^d y_i^B > 0$. Then p is calibrated on $A|B$ for $\{i : y_i^B = 1\}$ (i.e. for the set of steps where B occurs, see first line below) if and only if it satisfies $p(A|B) = \frac{p(A\cap B)}{p(B)}$:

$$
\begin{aligned}
\sum_{i:y_i^B=1} p(A|B) &= \sum_{i:y_i^B=1} y_i^A \\
\sum_{i:y_i^B=1} p(A|B) &= \sum_{i:y_i^B=1} y_i^{A\cap B} && \text{(since } y_i^A = y_i^{A\cap B} \text{ when } y_i^B = 1\text{)} \\
\sum_{i:y_i^B=1} p(A|B) &= \sum_{i=1}^d y_i^{A\cap B} && \text{(since } y_i^{A\cap B} = 0 \text{ when } y_i^B = 0\text{)} \\
\sum_{i:y_i^B=1} p(A|B) &= \sum_{i=1}^d p(A \cap B) && \text{(by calibration of } p \text{ on } A \cap B\text{)} \\
p(B) \cdot \sum_{i=1}^d p(A|B) &= \sum_{i=1}^d p(A \cap B) && \text{(by calibration of } p \text{ on } B\text{)} \\
p(A|B) &= \frac{p(A \cap B)}{p(B)}.
\end{aligned}
$$

Summing up, the probability calculus can be seen as a sound and complete system for generating calibrated predictions on certain sets from calibrated predictions on related sets. While this does not settle the question whether all predictors need to follow the probability calculus (i.e. that they are probability measures in their second argument), it does provide a *pro tanto* reason.

4.5 A UNIFIED INTERPRETATION OF PROBABILITY

> Norms of belief are as remote from empirical claims about nature as is Hume's simpler subjectivism. Propensity theories of probability propose a physical property that cannot be recorded and does not necessitate or preclude any occurrence. [...] any limiting-frequency claim is consistent with any claim about any finite collection of events.
>
> — Clark Glymour (2001)

Russell's famous dictum that 'probability is the most important concept in modern science, especially as nobody has the slightest notion what it means' (cited by Bell (1945, p. 582)) is almost a century old; but while there have certainly been many new developments, a satisfying interpretation is still lacking. The purpose of this section is to argue that his lacuna can be filled by the perspective on probabilities put forward in the present paper. To make this argument, we now specifically relate our account to the literature on interpretations of probability. In the most authoritative up-to-date treatment of the subject, Hájek (2019) asks 'what do we want from our interpretations *of probability*, specifically?' (original emphasis) and then answers by suggesting a list of desiderata (drawing on Salmon (1966)). Some of these we have already covered above: Our account satisfies 'non-triviality' (not just zero and one) and 'admissibility with respect to this or that axiomatization' (motivating the axioms of the probability calculus); it also illuminates 'ampliative inferences' in the sense that it allows to reason about the justification of probabilistic statements based on other probabilistic statements (Section 4.4). We now, in turn, discuss Hájek's remaining desiderata: the applicability to science, rational belief, frequencies, and rational decision making.

4.5.1 *Relation to science and the question of (in-)determinism*

> During the nineteenth century it became possible to see that the world might be regular and yet not subject to universal laws of nature. A space was cleared for chance.
>
> — Ian Hacking (1990)

Assessing the applicability to science means checking the compatibility with scientific practice and current scientific theories. In particular, the question of whether the universe is deterministic is sometimes taken to bear directly on

how we should think about probability. For example, David Lewis (1980, p. 120) thought that objective or physical probabilities rely on indeterminism. Karl Popper (1959) also proposed the propensity account of probability (according to which probabilities are physical properties) in the context of Quantum Mechanics (QM). We should briefly note that QM does not require the world to be indeterministic, given that different, empirically indistinguishable interpretations of QM disagree on this question (not even getting into the question of scientific realism). So there is certainly no need to presuppose this. But even if QM came with some notion of true probabilities, it is unclear that they would be of any relevance to the probabilities we deal with day-to-day, for two reasons: First, QM probabilities need to be described by a more general theory of probability than that axiomatised by Kolmogorov (Streater, 2000). Second and more importantly, even for a coin flip, we would never have access to the true QM-based probabilities: We are neither able to determine the initial conditions, that is, the complete wave function, nor to take into account the extremely high number of occurring quantum interactions.[11] The resulting values may, thus, vastly differ from any predictions we are able to make—which, as we have shown, are still useful. In sum, we do not see compelling reasons to suppose either determinism or indeterminism, nor to think that indeterminism at the level of QM would contribute much to probabilistic reasoning. It is therefore a strength of our account that, showing how probabilities can be constructed and used, it remains agnostic regarding the question of determinism.

Indeed, the intuition that probabilities are objective may depend less on QM and more on the often strong interpersonal agreements about 'correct' prediction methods e.g. for gambling. In the words of Michael Strevens (2006, p. 31), probabilities in such settings 'have attained a certain kind of stability under the impact of additional information. This stability gives them the appearance of objectivity, hence of reality, hence of physicality'. We argued in Section 4.2.2 that this appearance is misleading, as illustrated by the roulette story of Thorp and Shannon. In line with this, the 'erosion of determinism' indeed did not follow the advent of QM but of higher-level statistical regularities discovered during the previous century, as captured in Hacking's epigraph above. It is also important to note that the higher-level sciences depend heavily on abstractions such as the ones that feature in our notion of prediction methods. Abstractions have been observed to be both part of scientists' tacit knowledge (Polanyi, 1958), especially the 'ability to recognize a given situation as like some and unlike others' (Kuhn, 2012, p. 195), and a substantial part of conscious scientific work (Danks, 2015; Potochnik, 2017)—yet they, arguably, still remain an under-explored topic.

11 This would require a QM version of Laplace's demon.

4.5.2 *Relation to (rational) degrees of belief*

> Chances are degrees of belief [...]; not those of any actual person, but in a simplified system to which those of actual people, especially the speaker, in part approximate.
>
> — Frank P. Ramsey (1928/1931)

While probabilities are often said to have a direct connection to degrees of belief, we argue that the notion of probability does not depend on degrees of belief: prediction methods can be used in a purely mechanical way to get desirable outcomes on aggregate, without an entity involved that is commonly held to have beliefs. A simple machine that takes measurements and uses a predictor could make decisions based on probabilities without it being plausible to ascribe to it beliefs that come in degrees. However, probabilities can also be used to *describe degrees of belief* under uncertainty. In most cases, it is difficult to pin down a particular prediction method, especially as human predictions tend to be qualitative. But humans also take only specific information into account and exploit regularities such as symmetries or stable relative frequencies. In some cases, the gap between human reasoning and quantitative prediction methods can become fairly small—it appears that some people, modestly described as 'super-forecasters', are particularly good at making calibrated quantitative predictions (Mellers et al., 2015). The notion of prediction methods can also shed light on imprecise notions of (subjective) uncertainty. Particularly vague degrees of belief or disagreements between different methods can be represented by imprecise predictions (Appendix 4.8.1.3)[12] while Knightian uncertainty (Knight, 1921) corresponds to the absence of a trusted prediction method. In this sense, Section 4.2 also tells us an idealised story of human reasoning: Consciously or not, humans often implement something close to prediction methods—in that sense, probabilities can model human degrees of belief, a view also expressed in Ramsey's epigraph.

While probabilities should not be taken as actual degrees of belief, we can also *model consistent decision-making* by humans or machines as if they had certain degrees of belief: Savage (1972) famously showed that actions which follow certain consistency criteria can be viewed as maximising expected utility for implicit utility functions and probabilities. Expected utility maximisation (EUM) is, thus, often taken to be an approximation of human decision-making, that is,

12 Such interval probabilities were also considered by De Finetti and Savage (1962) as models of degrees of belief in 'the case of a number of decision-makers who have to make a collective decision, and, second, the case of a single individual who experiences a 'kind of personality dissociation'' (Feduzi, Runde and Zappia, 2012, p. 348).

as a descriptive theory: We tend to make decisions such that good outcomes seem more likely to us. There are, of course, considerable caveats. The most crucial ones are arguably diminishing marginal utility and risk aversion, already highlighted by Ramsey (1926/1931, p. 172) and analysed e.g. in (Wakker, 1994). In Ramsey's words, EUM as a modelling tool is an 'artificial system of psychology, which like Newtonian mechanics can, I think, still be profitably used even though it is known to be false' (Ramsey, 1926/1931, p. 173). So, we can connect degrees of belief to probabilities, by modelling reasoning and decision-making through prediction methods and Savage-style decision theory–but stop short of equating them.

There is, of course, some flexibility in deciding what to use the word 'probability' for. One may want to use it for the assignments in Savage-style models, that is, for implicit degrees of belief that can be assigned whenever some agent acts consistently. It seems, however, more consistent with everyday use to reserve it for the predictions themselves, for weather predictions and coin flips, and to say that we can model consistent decision as if they follow EUM under certain probabilities. As shown in Section 4.4.2, we can get calibrated predictions from calibrated predictions of related events using the probability calculus. This provides a *pro tanto* reason for considering the probability axioms to constitute constraints on rationality. We emphasise again that we do not put forward a normative theory of rational belief or action here, but a descriptive perspective on probability that can illuminate its perceived connections to rationality.

4.5.3 *Relation to frequencies and the reference class problem*

> If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the *problem of the reference class*.
>
> — Hans Reichenbach (1949)

Although our perspective implies that probabilities are constructed, it also explains their strong connection to observed relative frequencies. Indeed, if we restricted calibration to sets of equal probability (as calibration is sometimes understood), the relationship would be even closer: Then, the calibration condition would be equivalent to the definition of probability in finitary frequentism:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} y_i \qquad \Leftrightarrow \qquad p_i = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{4.22}$$

Instead of taking this as a definition, we think it more adequate to see it as a special case of our main quality criterion. Not just because such finitary definitions are problematic (Hájek, 1996),[13] but also because it would imply too narrow an evaluation criterion.

The ties between frequentism and our account become particularly clear in reference to the so-called reference class problem. Its metaphysical version is a problem for objectivist theories like frequentism that claim a unique true probability for each event (Hájek, 2007). The epistemic version concerns the question of how a reference class should be chosen for a given event, considering that different choices would lead to different probabilities. This has led Hájek, for example, to argue that conditional probabilities are actually primitive, as probabilities are always conditional on a certain conceptualisation of events. While our predictors do resemble them, they are not strictly speaking conditional probabilities, as $\mathcal{X}$ is just an arbitrary set without well-defined probabilities; as we show in Section 4.4.2, it makes more sense to consider conditional probabilities to further depend on a particular (perhaps implicit) abstraction $x \in \mathcal{X}$ (see the implicit joint 'conditioning' in Appendix 4.8.2). Hájek (2007) also supposed that, rather than a marginal probability, '[v]arious frequentists could tell us the conditional probability that John Smith will live to age 61, *given* that he is a consumptive Englishman aged 50' (ibid., p. 582, original emphasis). However, there can be different mortality tables resulting in different ratios—more generally, the choice of abstraction does not yet fix the prediction. This aspect is clear for the prediction method perspective, as different methods may use the same scheme of abstraction but different predictors, relying e.g. on different mortality tables.[14] For these reasons, we agree with Freedman (1997, p. 23) that 'probability is a subtler idea than relative frequency'. We would also argue that the reference class problem is not actually a problem. Different prediction methods may be calibrated on different sets, so one can choose a prediction method that promises calibration on sets of interest. This relates to to the more general idea of the 'goal-dependence in scientific ontology' (Danks, 2015).

13 Glymour (2001) argues for what he calls an instrumentalist and approximate version of finite frequentism which takes probabilities to be descriptions of frequencies rather than defining the former through the latter. While this is not too far from our somewhat pragmatic approach in spirit, his focus is more on the description of populations through distributions and he rejects the relevance of decision theory.

14 In Machine Learning, the phenomenon that prediction methods can give different predictions despite using the same abstractions *and* using the same data *and* achieving the same average loss is known under the name of 'predictive/model multiplicity' (see e.g. (Black, Raghavan and Barocas, 2022; Breiman, 2001)).

4.5.4 *Relation to rational decision-making and individual predictions*

> Legends of prediction are common throughout the whole Household of Man. God speaks, spirits speak, computers speak. Oracular ambiguity or statistical probability provides loopholes, and discrepancies are expunged by Faith.
>
> — Ursula K. Le Guin (1969): *The Left Hand of Darkness*

Our discussions have focused on sets of predictions rather than individual ones. This is not a coincidence, as we take probabilities to not be free-floating numbers but to rely on prediction methods which are useful when sets of predictions are calibrated.[15] But what exactly is the relation between a prediction and the corresponding event? And can we evaluate the quality of a single non-trivial prediction? That is, is a prediction of 0.6 better than a prediction of 0.4 if the predicted event occurs? What should we do if we only get a single prediction (for some level of utility)?

The first three questions all relate to the dependence of a prediction on the method that generated it. It is important to emphasise that the probability is not a property of the event, as it is constructed and depends on the choices of both the abstraction and the predictor. As discussed in Sections 4.2.2 and 4.5.1, gambling setups only appear to have objective probabilities because of their relative stability under additional information We also mentioned the example of Angelina Jolie, who stated 'My doctors estimated that I had an 87 per cent risk of breast cancer', with the number presumably coming from statistical data about women with a particular genetic mutation. Dawid (2017) asks, 'Was Angelina (or her doctors) right to interpret it as her own individual risk?' (p. 3456). On our account, they were—with the qualification that this risk is model-dependent and constructed rather than objective and discovered—as is any other probability. Which prediction method is most useful depends on which sets we want to be calibrated on (although the perfect binary predictor is always optimal). In hindsight, we can usually say which prediction would have been good or correct. The validation of prediction methods cannot, however, be thus reduced to comparisons between individual predictions, not even with proper scoring rules (Gneiting and Raftery, 2007). They do allow us to put a number on our intuition that 0.6 is somehow a better prediction than 0.4 if the predicted event occurs; but so does any notion of calibration error (as the ℓ_1 loss deployed in Appendix 4.8.1.1). After all, proper scoring rules are meant to be 'appropriate

15 Of course, singletons can be (approximately) calibrated when predictions are (approximately) 0 or 1.

for evaluating and comparing forecasters who *repeatedly present their predictions*' (DeGroot and Fienberg, 1983, p. 12, emphasis added).

The fourth and last question about acting on single predictions is related but more complex. In general, we suggest that policies rather than single actions should be the subject of justification and evaluation. An ex-post evaluation of a decision would ignore the prediction and just consider whether an alternative decision would have been better in hindsight—this is not particularly helpful. Instead, what is familiar also from legal and ethical reasoning (especially deontological, but even rule-consequentialist), is to judge decisions by the reasons or maxims that they were based on.[16] For example, we showed that maximising expected utility is a good policy if we can assume calibration on sets of equal utility and wish to maximise cumulative utility (Section 4.3.1). As noted before, being calibrated for all possible combinations of decisions would require perfect discrimination. In the umbrella example of Section 4.3.3, we showed that the calibration criterion can be more benign when comparing a more restricted set of sensible policies. But the problem is more difficult e.g. when we only have a few predictions for particularly grave events: If we only make a few high-stakes decisions, such as a choice of treatment for breast cancer (where one may even argue that the concept of numerical utility breaks down), it seems too big of an assumption to hope for calibration on such a small set. That being said, the combinatorial reasoning from Section 4.4.1 also extends to single predictions: Good calibration in the past gives *some* reason to believe in low calibration error on the single prediction, i.e. the more reason to believe in the event, the higher the prediction: The strength of this mathematically-grounded pro-tanto reason is monotonous in the number of events, so some (even if small) reason to believe will remain.[17] In general, for such situations, it may be more sensible to be risk-averse than in low-stakes settings where there are multiple events with comparable utility (cf. Buchak, 2013; Thoma, 2019).[18] A way to model this would be via imprecise calibration as explored in Appendix 4.8.1.3.

16 This has been stated in particularly succinct form by Maurice Merleau-Ponty (1955, p. 9): 'Il n'y a pas des décisions justes, il n'y a qu'une politique juste.'

17 I am grateful to Gunnar König for pushing me on this point.

18 This creates an asymmetry for the doctor-patient relationship, but also for algorithmic predictions, similar to the insurance setting (Fröhlich and Williamson, 2024a). C.S. Peirce (1878), in contrast, thought that when probabilistic reasoning is confronted with limited trials, 'logicality inexorably requires that our interests [...] must not stop at our own fate, but must embrace the whole community'.

4.6 COMPARISON WITH CONVENTIONAL INTERPRETATIONS

Hájek (2019) notes that '[e]ach interpretation that we have canvassed seems to capture some crucial insight into a concept of [probability], yet falls short of doing complete justice to this concept.' Any new satisfactory account of probability should, thus, be expected to make proponents of other accounts feel vindicated on some aspects that are particularly close to their hearts. We think that this is the case for our notion of probabilities as outputs of prediction methods aiming to predict numbers of occurring events. It is interesting to note, for example, that our predictors $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ resemble the confirmation function central to logical accounts of probability, such as that of Carnap (1950) or Keynes (1921) (with precursors as early as Leibniz, cf. (Hacking, 1975)). Our predictors, however, are neither objective relations nor relations between propositions—they are functions of abstractions in a set $\mathcal{X}$ and events in an algebra $\mathcal{A}$. In this section, we briefly survey a number of other prominent interpretations and highlight what we take to be the most interesting similarities and differences w.r.t. our account.

Bayesianism roughly posits that probability and its theory are concerned with degrees of belief and rationality constraints thereon. What we agree with is that probabilities are constructed and that it is misguided to search for true probabilities. However, we ground them in prediction methods rather than degrees of belief (Section 4.5.2) and highlight that these methods aim to track structure in sets of observations. This makes it possible to replace notions of internal cohesion or rationality with that of empirical calibration, and thereby a guide to decision-making that guarantees good outcomes. A Bayesian account that is particularly close to ours is that of Philip Dawid (2017).[19] On the one hand, his suggestion to arrive at 'probability forecast[s] by assessing the odds at which I would be willing to bet' (p. 3471) is clearly Bayesian in the tradition of (De Finetti, 1937). On the other hand, he also suggests to evaluate individual predictions on aggregate data via calibration—although its precise scope and relevance do not become entirely clear. In particular, it remains unclear why calibration on future data is important and on which (finite/infinite) sets it matters.[20] In comparison, our notion of prediction methods focuses on (potentially) inter-subjective models and the role of abstraction, which is decoupled from the events $A \in \mathcal{A}$ that we wish to be calibrated on. In a way, then, we posit a variant

19 Also the work of (Shafer and Vovk, 2019), which builds on Dawid's earlier work, is quite close to ours in spirit; but their focus is on testing probability forecasts, rather than how they are useful.

20 For example, his notion of H-based calibration seems to require calibration on all sets that cannot be further distinguished—which can amount to calibration on individual datapoints.

of Bayesianism without degrees of belief or betting and with a more concrete connection to the world, enabling not only the avoidance of sure loss in Dutch books but successful action in everyday life.

Hypothetical *frequentism* can be defined as the suggestion that 'the probability of an attribute A in a reference class B is the value the limiting relative frequency of occurrences of A within B would be if B were infinite' (Hájek, 2019). This captures the intuition of identifying probabilities with ratios in repeated trials. While this sounds very different to our account at first glance, we already discussed two similarities in Section 4.5.3: One is the dependence of individual probabilities on other events and on a choice of abstraction (via prediction methods, in our case), leading us to a generalisation of the reference class problem. Furthermore, equating probabilities with relative frequencies is a special case of our notion of calibration, which we consider for finite sets. We do reject the jump to declaring that probabilities themselves are 'out there' in any interesting sense. The finite frequentist account of Glymour (2001), mentioned in footnote 13, provides, in a sense, an intermediate account.

Karl Popper abandoned frequentism in favour of his *propensity* account because the former could not make sense of sequences with few trials. He thus proposed that frequentists should alter their theory by letting it 'say that admissible sequences must be either virtual or actual sequences which are *characterised by a set of generating conditions*—by a set of conditions whose repeated realisation produces the elements of the sequence' (Popper, 1959, p. 34, original emphasis). This is still an objectivist theory but dispenses with the reliance on infinite trials, instead invoking a new sort of mysterious property (especially in the case of a deterministic universe, which Popper did not seem to assume). We argued that relevant probabilities are independent of 'true' probabilities that may or may not be implied by Quantum Mechanics (Section 4.5.1). Propensity accounts often have a frequentist flavour, indirectly highlighting the importance of sets of events. It is interesting to note that, as the equivalence classes of generative conditions are idealisations (ignoring background conditions, cf. Section 4.2.2), they can be seen as abstractions made by prediction methods. However, propensities are typically thought to be physical rather than model-dependent, which is in stark contrast to our account—although the relevant literature sometimes also invites a reading of model-dependent propensities.

Another interesting interpretation of probability is the *best-systems account* of David Lewis (1994), which also posits objective chances: On this view, 'the chances are what the probabilistic laws of the best system say they are' (p. 480). 'The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. [...] If nature is kind, the best system will be

robustly best [...] It's a reasonable hope' (ibid., p. 478f). Now this account presupposes what may seem a tremendous kindness of nature as well as a perhaps weak notion of truth and objectivity—the latter fits well into Lewis' Humean view on laws of nature. What is interesting here about Lewis' account is that it resonates with the hope for a best level of predictive depth expressed by Dawid (2017, p. 3465)—in turn similar to the 'primary resolution' of Li and Meng (2021). Indeed, Dawid could be seen as linking the best-systems view on probability with our more pragmatic notion of model-based predictions. The clearest differences on the side of Lewis are the integration within a more global systematisation of the universe and the belief in objectivity, hinging on the existence of a privileged description. If there were an objectively best predictor and we assigned to it some notion of truth, these differences would blur.[21] However, this hope for or pretension of objectivity is also what Clark Glymour criticises in typical frequentist takes.

In line with our analysis, Glymour thinks that central problems with Bayesianism and frequentism lie, respectively, in the neglect of empirical claims and the unnecessary stipulation of objectively true probabilistic statements:

> The sometimes bitter debates between those who describe themselves as frequentists and those who describe themselves as subjective Bayesians has often turned on charges by the former that the latter abandon the "objectivity" of science and by the latter that the former dissemble about the "subjectivity" of their probability judgements. My belief is that, among statisticians anyway, the dispute often confuses content with justification. The "objectivity" of the frequentists is in the content of their probability judgements, which, while usually stated as about an unempirical probability, are often really vague empirical claims about finite frequencies. That sort of objectivity is genuinely lost in subjective Bayesian interpretations. The "subjectivity" kept hidden by frequentists is that there is often no explicit justification beyond their own opinion for aspects of their empirical claims. That subjectivity can be made entirely explicit without sacrificing the objective–that is empirical–content of frequency claims, and its recognition does not require, or even invite, recourse to subjective probability. Bayesian criticisms do address a confused and uncertain frequentist statistical practice, in which the point of making empirical claims is often forgotten or fudged. (Glymour, 2001, p. 299f)

21 This is perhaps not surprising given the subjectivism and pragmatism of Frank Ramsey, whom Lewis credits with a first formulation of a best-systems approach.

We have argued that our account avoids these problems by stating that probabilities are constructed rather than discovered while still taking their justification directly from empirical observations. Even more, we connect successful decision-making with empirical evaluation and assumptions about induction through a general notion of calibration, which has not been considered a central concept by any of the conventional accounts.

4.7 CONCLUSION

Relying on the notions of finite calibration and prediction methods, we have provided a more or less pragmatic account of probabilities and how they are useful for decision-making. We showed that if predictions satisfy an (often feasible) calibration criterion, then it is possible to predict the distribution of utilities that a given policy will yield. In particular, the sum of one's predicted utilities will match the actual cumulative utility, which can provide a rationale for expected utility maximisation. A central element of our account is the semi-formal notion of prediction methods that construct abstractions of given situations and feed them to a model. Arguably, the novelty here consists less in the consideration of predictions than in the connections drawn to abstractions and to successful decision-making via calibration in very general terms. Indeed, we argue that this perspective elucidates the relationship between gambling odds and rain predictions, uniting the alleged two faces of probability by tying together abstraction, forecasting, probability theory, and empirically successful decision-making.

The understanding of probability also has important implications for data science. For example, it underscores the relevance of evaluating calibration beyond sets of equal predictions, as already explored by Dawid (2017) and Höltgen and Williamson (2023). Furthermore, it underscores that one should be wary of the often-invoked concept of a 'true distribution' from which one can 'sample' once one steps outside of the casino or other highly controlled settings. Given the centrality of probability for causality, many considerations also spill over to the latter; indeed, we take a deeper dive into causal inference from this perspective in concurrent work, also highlighting the role of calibration (Höltgen and Williamson, 2026a). It has been observed that algorithmic predictions based on machine learning tend to convey an air of authority and objectivity, as the many choices involved in data collection (abstraction scheme) and model tuning (choice of predictor) often remain beneath the surface (Moss, 2021). This is particularly relevant for the justification of predictions that inform decisions about people (Höltgen and Williamson, 2026b). Our work highlights that probabilities,

e.g. of finding a job, are not properties of people; instead, they depend on the selected abstraction and model, which, in turn, depends on data about other people. Hence also our answer to Cynthia Dwork's question from the introduction: There is no ideal algorithm, as there are no true probabilities to uncover, and different algorithms can be better suited for different goals. While we hope that this work helps to sharpen the view on probability, a sea of open questions still calls for further exploration.

4.8 APPENDIX

4.8.1 *Generalising Proposition 4.5*

4.8.1.1 *Approximate calibration*

Here, we generalise Proposition 4.5 to only require approximate calibration—we give a bound on how large the calibration error on each set of equal utility can be in order to keep the difference between predicted and cumulative utility below some $\epsilon > 0$.

Proposition 4.9 (Predicting cumulative utility: Approximate calibration)**.** *We assume the same setting as in Proposition 4.5 except that we now require all utilities to be positive—one may otherwise simply shift the values to a positive domain. If we then replace condition (4.3) with the assumption that* $\forall u' \in \mathcal{U}$,

$$\left| \sum_{i:u_i(0)=u'} y_i - \sum_{i:u_i(0)=u'} p_i \right| \leqslant \frac{\epsilon}{2 \cdot u' \cdot |\mathcal{U}|} \tag{4.23}$$

and the same for $u_i(1)$*, then*

$$\left| \sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] - \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)] \right| \leqslant \epsilon. \tag{4.24}$$

Proof.

$$\left|\sum_{i=1}^{d}[y_i u_i(1)+(1-y_i)u_i(0)]-\sum_{i=1}^{d}[p_i u_i(1)+(1-p_i)u_i(0)]\right|$$

$$=\left|\sum_{u'\in\mathcal{U}}\left(\left(\sum_{i:u_i(1)=u'}y_i-\sum_{i:u_i(1)=u'}p_i\right)\right.\right. \tag{4.25}$$

$$\left.\left.+\left(\sum_{i:u_i(0)=u'}p_i-\sum_{i:u_i(0)=u'}y_i\right)\right)\cdot u'\right| \tag{4.26}$$

$$\leqslant\sum_{u'\in\mathcal{U}}\left(\left|\sum_{i:u_i(1)=u'}y_i-\sum_{i:u_i(1)=u'}p_i\right|\right. \tag{4.27}$$

$$\left.+\left|\sum_{i:u_i(0)=u'}p_i-\sum_{i:u_i(0)=u'}y_i\right|\right)\cdot u' \tag{4.28}$$

$$\leqslant\sum_{u'\in\mathcal{U}}\left(\frac{\epsilon}{2\cdot u'\cdot|\mathcal{U}|}+\frac{\epsilon}{2\cdot u'\cdot|\mathcal{U}|}\right)\cdot u' \tag{4.29}$$

$$=\epsilon \tag{4.30}$$

□

While we use a symmetric ℓ^1 loss here, it may be interesting to also look into other measures of error. For example, for settings where under-prediction and over-prediction are valued differently, it may be instructive to look into asymmetric error functions.

4.8.1.2 *Approximate utility level sets*

Here, we generalise Proposition 4.5 to only require calibration on sets of approximately equal utility: For this, we divide the utility spectrum into bins of some size $\delta > 0$ and bound the resulting difference between predicted and cumulative utility by a term dependent on δ and the number of predictions d.

Proposition 4.10 (Predicting cumulative utility: Approximate utility)**.**
We assume the same setting as in Proposition 4.5 except that we now require all utilities to be positive—otherwise, one may simply shift the values to a positive domain. We partition the interval of relevant utilities from the lowest $u_i(a)$ to the highest $u_i(a)$ with $i \in \{1,...,d\}, a \in \{0,1\}$ into bins $B_1,...,B_m$ of size $\leqslant \delta$. If we then replace condition (4.3) with the assumption that $\forall k \in \{1,...,m\}$,

$$\sum_{i:u_i(0)\in B_k}y_i=\sum_{i:u_i(0)\in B_k}p_i \quad \textit{and} \quad \sum_{i:u_i(1)\in B_k}y_i=\sum_{i:u_i(1)\in B_k}p_i \tag{4.31}$$

then

$$\left|\sum_{i=1}^{d}[y_i u_i(1)+(1-y_i)u_i(0)]-\sum_{i=1}^{d}[p_i u_i(1)+(1-p_i)u_i(0)]\right| \leqslant \delta\cdot d. \quad (4.32)$$

Proof.
The maximal mismatch occurs when for each bin B_k and each $a \in \{0,1\}$, one half of the $\{i : u_i(a) \in B_k\}$, we have $(y_i - p_i) = 1$ and $u_i(a) = m_k + \delta/2$ whereas for the other half, $(y_i - p_i) = -1$ and $u_i(a) = m_k - \delta/2$, with m_k denoting the midpoint of B_k. This gives

$$\forall k \in \{1,...,m\}, a \in \{0,1\}: \quad \left|\sum_{i:u_i(a)\in B_k}(y_i-p_i)\cdot u_i(a)\right| \leqslant \left|\sum_{i:u_i(a)\in B_k}\delta/2\right| \quad (4.33)$$

$$= b_k^a \cdot \delta/2 \quad (4.34)$$

where $b_k^a := |\{1 \leqslant i \leqslant d \mid u_i(a) \in B_k\}|$. Therefore,

$$\left|\sum_{i=1}^{d}[y_i u_i(1)+(1-y_i)u_i(0)]-\sum_{i=1}^{d}[p_i u_i(1)+(1-p_i)u_i(0)]\right| \quad (4.35)$$

$$\leqslant \left|\sum_{i=1}^{d}[y_i\cdot u_i(1)-p_i\cdot u_i(1)]\right| + \left|\sum_{i=1}^{d}[(1-y_i)\cdot u_i(0)-(1-p_i)\cdot u_i(0)]\right| \quad (4.36)$$

$$= \left|\sum_{i=1}^{d}[(y_i-p_i)\cdot u_i(1)]\right| + \left|\sum_{i=1}^{d}[(y_i-p_i)\cdot u_i(0)]\right| \quad (4.37)$$

$$\leqslant \sum_{k=1}^{m}\left(\left|\sum_{i:u_i(1)\in B_k}(y_i-p_i)\cdot u_i(1)\right| + \left|\sum_{i:u_i(0)\in B_k}(y_i-p_i)\cdot u_i(0)\right|\right) \quad (4.38)$$

$$\leqslant \sum_{k=1}^{m}\left(b_k^1\cdot\delta/2 + b_k^0\cdot\delta/2\right) \quad (4.39)$$

$$= \delta\cdot d \quad (4.40)$$

□

4.8.1.3 *Imprecise calibration*

We now consider imprecise forecasts which give interval predictions $[a,b] \subset \mathbb{R}$ and represent them as tuples $p^* = (\underline{p}, \bar{p}) \in \mathbb{R}^2$ of the lower and upper probability. This allows for a weaker calibration criterion where the number of occurring events need not exactly match the sum of predictions, but should lie between the sum of the lower and the sum of the higher predictions.

Definition 4.11 (Imprecise calibration)**.**
Imprecise predictions $p_1^*, ..., p_d^*$ *are said to be* imprecisely calibrated *for observations* $y_1, ..., y_d \in \{0, 1\}$ *if they satisfy*

$$\sum_{i=1}^{d} \underline{p}_i \leqslant \sum_{i=1}^{d} y_i \quad \textit{and} \quad \sum_{i=1}^{d} \bar{p}_i \geqslant \sum_{i=1}^{d} y_i. \tag{4.41}$$

Note that for our definition, the vacuous forecast that always predicts $(0, 1)$ is always imprecisely calibrated.[22] One could also apply the criterion of imprecise calibration to a set of precise predictions, by simply converting every precise prediction p_i into an imprecise forecast $[p_i - \epsilon, p_i + \epsilon]$ for some ϵ—this epsilon may also monotonically decrease in d to account for lower variance on larger sets.

Proposition 4.12 (Predicting cumulative utility, imprecise version)**.**
Let there be d *imprecise predictions* $p_i^* \in \mathbb{R}^2$ *for binary outcomes* $y_i \in \{0, 1\}$, $i \in \{1, ..., d\}$ *with utility functions* $u_i : \{0, 1\} \to \mathbb{R}$ *and assume that the predictions are imprecisely calibrated on sets of equal utility* $u_i(0)$ *and on sets of equal utility* $u_i(1)$ *(formalised in (4.44) below).*
Then I can correctly predict a range for my cumulative utility (LHS) via

$$\sum_{i=1}^{d} [y_i u_i(1) + (1 - y_i) u_i(0)] > \sum_{i=1}^{d} \left[\underline{p}_i u_i(1) + (1 - \underline{p}_i) u_i(0)\right] \tag{4.42}$$

and

$$\sum_{i=1}^{d} [y_i u_i(1) + (1 - y_i) u_i(0)] < \sum_{i=1}^{d} [\bar{p}_i u_i(1) + (1 - \bar{p}_i) u_i(0)] \tag{4.43}$$

The proof is analogous to that of Proposition 4.5, now with the calibration assumptions

$$\begin{aligned} \sum_{i:u_i(1)=u'} y_i > \sum_{i:u_i(1)=u'} \underline{p}_i \quad &\text{and} \quad \sum_{i:u_i(0)=u'} y_i > \sum_{i:u_i(0)=u'} \underline{p}_i, \\ \sum_{i:u_i(1)=u'} y_i < \sum_{i:u_i(1)=u'} \bar{p}_i \quad &\text{and} \quad \sum_{i:u_i(0)=u'} y_i < \sum_{i:u_i(0)=u'} \bar{p}_i. \end{aligned} \tag{4.44}$$

This allows people to not only optimise their utility but to also take risk-averse or risk-seeking inclinations into account—selecting policies not based on the expected exact cumulative utility but on e.g. the lowest or highest estimation of it. Here, we can see an analogy between the move from deterministic to

22 Similar issues are discussed in the literature on imprecise probability (Walley, 1991), particularly for Brier-style scoring rules (Seidenfeld, Schervish and Kadane, 2012) and randomness (De Cooman and De Bock, 2022, Prop. 9).

probabilistic and the move from precise to imprecise predictions: The former allows people to take their (cardinal) preferences into account (Section 4.3.3), whereas the latter allows them to take their risk aversion into account. If we know that we will be calibrated, risk aversion does not make much sense. Cases where we are less sure of it can be represented by an assumption of imprecise calibration. Note that this notion of risk-aversion also captures unwillingness to bet, for decisions between the utility function of a bet and the constant zero utility function with $u(0) = u(1) = 0$.

4.8.2 *Conditional probabilities, generalised*

We here generalise the analysis of *conditional probabilities* from Section 4.4.2. Consider a predictor $p : \mathcal{X} \times \mathcal{A} \to [0,1]$, events $A_1, ..., A_d, B \in \mathcal{A}$, and inputs $x_1, ..., x_d \in \mathcal{X}$. We assume $p(x_1, B) = ... = p(x_d, B)$ and that p is calibrated on $\{(x_i, A_i \cap B) : 1 \leqslant i \leqslant d\}$ and $\{(x_i, B) : 1 \leqslant i \leqslant d\}$, that is,

$$\sum_{i=1}^{d} p(x_i, A_i \cap B) = \sum_{i=1}^{d} y_i^{A_i \cap B} \tag{4.45}$$

and

$$p(x_1, B) = \frac{1}{d} \sum_{i=1}^{d} y_i^{B} > 0. \tag{4.46}$$

Then p is calibrated on $\{(x_i, A_i|B) : y_i^B = 1\}$ (i.e. for the set of steps where B occurs) if and only if it satisfies

$$\sum_{i=1}^{d} p(x_i, A_i|B) = \sum_{i=1}^{d} \frac{p(x_i, A_i \cap B)}{p(x_i, B)}, \tag{4.47}$$

as we derive below. In particular, a sufficient condition is

$$p(x_i, A_i|B) = \frac{p(x_i, A_i \cap B)}{p(x_i, B)}. \tag{4.48}$$

Now consider the special case where $A_i = ... = A_d =: A$ and $x_i = ... = x_d =: x$. Here, the familiar definition of conditional probabilities

$$p(x, A|B) = \frac{p(x, A \cap B)}{p(x, B)} \tag{4.49}$$

is necessary and sufficient for p to be calibrated for predictions of $A|B$ based on inputs x on the set of steps where B occurs. That is, making predictions for $A|B$ rather than A allows us to be calibrated on the set where B occurs (which predictions for A would usually not be).

Now the promised derivation of the characterisation (4.47):

$$\begin{aligned}
\sum_{i:y_i^B=1} p(x_i, A_i|B) &= \sum_{i:y_i^B=1} y_i^{A_i|B} \\
\sum_{i:y_i^B=1} p(x_i, A_i|B) &= \sum_{i:y_i^B=1} y_i^{A_i\cap B} && (*) \\
\sum_{i:y_i^B=1} p(x_i, A_i|B) &= \sum_{i=1}^{d} y_i^{A_i\cap B} && (\dagger) \\
\sum_{i:y_i^B=1} p(x_i, A_i|B) &= \sum_{i=1}^{d} p(x_i, A_i \cap B) && \text{(by (4.45))} \\
p(x_1, B) \cdot \sum_{i=1}^{d} \cdot p(x_i, A_i|B) &= \sum_{i=1}^{d} p(x_i, A_i \cap B) && \text{(by (4.46))} \\
\sum_{i=1}^{d} p(x_i, A_i|B) &= \sum_{i=1}^{d} \frac{p(x_i, A_i \cap B)}{p(x_i, B)};
\end{aligned}$$

whee $(*)$ follows from $y_i^{A_i|B} = y_i^{A_i} = y_i^{A_i\cap B}$ when $y_i^B = 1$ and $(\dagger)$ follows from $y_i^{A_i\cap B} = 0$ when $y_i^B = 0$.

ACKNOWLEDGEMENTS

For helpful feedback on previous versions, I would like to thank Ben Jantzen, Bob Williamson, Elisa Nguyen, Jannik Thümmel, Kate Vredenburgh, Konstantin Genin, Rabanus Derr, and Timo Freiesleben.

This work was funded by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. I also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for their support.

Part III

ALGORITHMIC PREDICTIONS IN SOCIETY

5

THE COSTS OF PRETENDING THAT THERE ARE DATA-GENERATING PROBABILITY DISTRIBUTIONS IN THE SOCIAL WORLD

The impersonal, collective character is thus the product *rather than the* producer *of the infinitely numerous individual characters;*

— Gabriel Tarde, *On Communication and Social Influence*

AUTHOR CONTRIBUTIONS

While the idea for the paper *per se* is mostly due to Benedikt Höltgen, the general scepticism of data-generating distribution as well as some of the ideas in the paper are strongly influenced by Robert Williamson. The project, including conceptualisation, theoretical results, and empirical analysis, was conducted by Benedikt Höltgen under the supervision of Robert Williamson. The writing is due to Benedikt Höltgen, with valuable feedback from Robert Williamson on multiple occasions.

ABSTRACT

Machine Learning research, including work promoting fair or equitable algorithms, often relies on the concept of a data-generating probability distribution. The standard presumption is that since data points are 'sampled from' such a distribution, one can learn from observed data about this distribution and, thus, predict future data points which are also drawn from it. We argue, however, that such true probability distributions do not exist and that the rhetoric around them is harmful in social settings. We show that alternative frameworks focusing directly on relevant populations rather than abstract distributions are available and leave classical learning theory almost unchanged. Furthermore, we argue that the assumption of true probabilities or data-generating distributions can be misleading and obscure both the choices made and the goals

pursued in machine learning practice. Based on these considerations, we suggest to avoid the assumption of data-generating probability distributions in the social world.

5.1 INTRODUCTION

We have grown so accustomed to probabilistic statements that we hardly question what they entail. Particularly in Machine Learning (ML) and Statistics, probabilities and probability distributions are ubiquitous. Students are required to learn the basics of probability theory since virtually all theoretical work in ML is embedded in a probabilistic framework. This makes sense since most of ML is concerned with prediction, and predictions contain uncertainty—which can often be captured through probability theory. However, probabilities play a central role not only as predictions but also as true distributions that our data is sampled from; this is the general theoretical framework that we use to think about machine learning, for example to derive error guarantees (Vapnik, 1982). But while we can do a lot of maths with probability theory, we still lack a reliable understanding of when and why our probabilistic models and frameworks actually work in practice.[1] The purpose of this paper is to argue that, at least in social settings, we should avoid the true data-generating distribution (Gen-D) framework that is ubiquitous in Machine Learning as well as Statistics. This applies particularly to work promoting fair or equitable algorithms.

It is a very common assumption in all areas of ML to say that an observed set of datapoints is sampled from some true joint distribution. This allows us to mathematically capture the (more or less stable) patterns we see in the world, and to formalise learning. This is a particular choice of a mathematical model; to avoid confusion with the specific 'models' trained on concrete data (such as neural networks), we will call it the Gen-D *framework* here. As researchers, we often make statements like 'we assume that our datapoints $\{(x_i, y_i)\}_{i=1}^N$ are sampled independently and identically distributed (i.i.d.) from the true distribution over $\mathcal{X} \times \mathcal{Y}$', even though we know that this is an idealisation. One may seek consolation in saying that all models are wrong anyway—but it is important to know in which ways they are wrong, in order to anticipate when they fail. The still common (since easy to work with) i.i.d. assumption is increasingly criticized in the community; however, the suggested adjustments typically only address either of the i's, especially the 'identical', via frameworks for distribution shift or data corruption. In contrast, we already see a more fundamental

1 This can serve as an illustration of John von Neumann's point that 'if people do not believe that mathematics is simple, it is only because they do not realize how complicated life is' (Alt, 1972).

problem with saying that our data is somehow sampled from some abstract true distribution.[2] The Gen-D assumption often also motivates the suggestion that the goal of learning is to approximate this distribution in terms of the 'Bayes-optimal predictor' $P(Y|X)$. And for particularly stable settings, such as gambling devices or pseudo-random number generators, data-generating distributions can indeed be a useful idealisation. In this paper, however, we argue that in social settings, when making predictions about people, we should not assume the existence of a true distribution, even as an idealisation.

We start, in Section 5.2, by arguing that true generative distributions do not exist, even if, within the limits of stable problem settings, they can be useful modelling tools. In Section 5.3, we then debunk their perceived raison d'être, namely that they are necessary to model the stabilities which we see in practice and, in particular, learning and generalisation; we show that one can modify classical learning theory so that it works with finite populations. In Section 5.4, we show how the Gen-D framework obscures how Machine Learning works and what it can provide. In particular, it may suggest that model-building is approximation rather than construction, as it was originally conceived. This has important implications, for example, on the interpretation of theoretical trade-offs between fairness and accuracy (Corbett-Davies et al., 2017; Hardt, Price and Srebro, 2016; Menon and Williamson, 2018), and their relation to empirical observations about model multiplicity and the possibility of less discriminative algorithms (Black et al., 2024a; Cooper et al., 2024; Höltgen and Oliver, 2025). We also argue that the Gen-D framework conceals the modelling choice of selecting an input space due to its framing of the learning problem, which can be an obstacle both to further research and to open discussion in applications. In Section 5.5, we highlight a sometimes underappreciated implication of our foregoing discussions, namely, that ML models are far from objective, as they depend on many design choices. Section 5.6 concludes.

5.2 GEN-DS DO NOT REALLY EXIST

In this section, we argue that there are no true probabilities and, by implication, no true generative distributions in ML. Still, people tend to have particularly strong intuitions in favour of objective probabilities in gambling settings such as coin flips or dice throws: there do seem to be 'correct' probabilities for such games of chance—but this is misleading. As put by Michael Strevens, probab-

2 A line of work that does not rely on this assumption is conformal prediction (Shafer and Vovk, 2008), although its reliance on exchangeability assumptions makes it vulnerable to similar critiques.

ilities in gambling settings 'have attained a certain kind of stability under the impact of additional information. This stability gives them the appearance of objectivity, hence of reality, hence of physicality' (Strevens, 2006, p. 31). Indeed, this stability is the very reason why games like roulette are used for gambling because there is no easy way to beat the casino.[3]

As Quantum Mechanics (QM) is often thought to prove the existence of true probabilities, we feel compelled to briefly dispute this. First, QM is consistent with determinism, e.g. via Bohmian mechanics (Detlef, Dürr and Teufel, 2009). Second, even if the world were indeterministic, this could mean that the QM-probability of one and the same coin turning up heads is above 50% in one situation and below 50% in another situation, due to minuscule differences in background conditions. But as we neither have access to the complete world state nor have the computational resources to compute the whole dynamics, the (non-)existence of such QM-probabilities is actually irrelevant for the interpretation of the macro-scale probabilities we deal with in everyday life and in virtually all settings where ML is applied. As Nancy Cartwright observes, the natural sciences and especially physics have focused on finding latent quantities that stand in stable relationships to each other, whereas in social settings we typically consider quantities that are of direct interest or easy to measure. And, as she puts it, 'to suppose that there really is some probability measure over [such quantities], you need a lot of good arguments' (Cartwright, 1999, p. 325). In sum, at least in social settings, it seems fair to say that probabilities are always constructed rather than discovered (Höltgen, 2026; Spiegelhalter, 2024).

But how about the data-generating distributions that are typically assumed in ML? One approach to try to justify the assumption of such a distribution could be to interpret it as the empirical distribution that *would* be observed if the whole population were to be queried. Take the example of predicting whether people in the US with a certain set of attributes will commit a crime within the next year. Consider the (in practice unobserved) empirical distribution that captures the whole population of the US for, say, 10 or 20 years. One could then regard an actually collected dataset (of, say, thousands of datapoints) as a sample from that empirical distribution (constructed from millions of datapoints). The two problems we will now point out are particularly clear for very fine-grained input spaces $\mathcal{X}$, that is, when we collect many attributes.

One problem is that the 'true' distribution depends very strongly on our implicit choices (beyond the choice of abstraction that we will discuss in Section 5.4.2). In particular, it depends on what we take to be the population de-

3 Though see (Thorp, 1998) for an anecdote of how Edward Thorp and Claude Shannon used hand-made devices to beat casinos at roulette.

scribed by the empirical distribution. For example, if there are few datapoints per $x \in \mathcal{X}$, then the true conditional probability $P(Y|X)$ depends a lot on the time frame that the distribution is supposed to cover: if there are only a few people with the characteristics x, then $P(Y|X = x)$ can strongly depend on whether we take into account 10 or 20 years, given that the observed ratios are typically not stable then. In the ML context, this can be seen, for example, in degradation of performance over time. In the example of predicting income based on US census data, the accuracy of models trained in one year continuously degrades every year (Ding et al., 2021), see also our results on individual states in Appendix 5.7.2.2 and other results across domains by Vela et al. (2022). We now want do draw attention to another aspect of this, also considering `ACS Income` (Ding et al., 2021): The predictions even of our best models, which are often considered close to Bayes-optimal as classifiers, significantly depend on the choice of population. In Figure 5.1, we show that around 10% of the predictions of tuned models change by at least 10%, depending on whether we consider the last year or the last four years This numbers apply to the four largest US states with very large datasets and are even higher for datasets of smaller states. The proportions are lower for some suboptimal coarser models, but even higher for comparably accurate but finer models (more details in Appendix 5.7.2.1). If the distribution depends on the year, how about the day or second a person is 'sampled'? Beyond time considerations, 'transfer from one state to another gives unpredictable results in terms of predictive accuracy and fairness criteria.' (Ding et al., 2021, p. 9) In general, making the choice of the considered population explicit is important for applying ML in social settings. This suggests a different understanding of generalisation, to which we will come back later.

Now take the example of predicting credit default, where there is likely *e.g.* a general trend that connects higher income with lower rates of credit default. Assume, then, for some $x_1 \in \mathcal{X}$ that includes very high income as well as otherwise favourable attributes, there are only three datapoints among the millions which fall into this description; assume further that two of these three people happened to default for whatever reasons. Would we want to say the 'true probability' of default for that combination of attributes is really $P(Y = 1|X = x_1) = 2/3$? Now the problem is not just that we *cannot* learn such a volatile conditional distribution, but that we do not *want* to: We aim to learn general patterns in the data that 'generalise',[4] not ratios between datapoints that

4 The common, unspecific talk of 'generalisation' leaves open its scope, i.e. the population to which it should generalise. As argued above, this typically goes unnoticed precisely because learning is always discussed in the context of a true distribution with undefined scope. If we want to predict well on unseen data, however, we need some reason to believe that it is similar to our training or validation data—but in what sense? This needs to be specified with reference to the model we

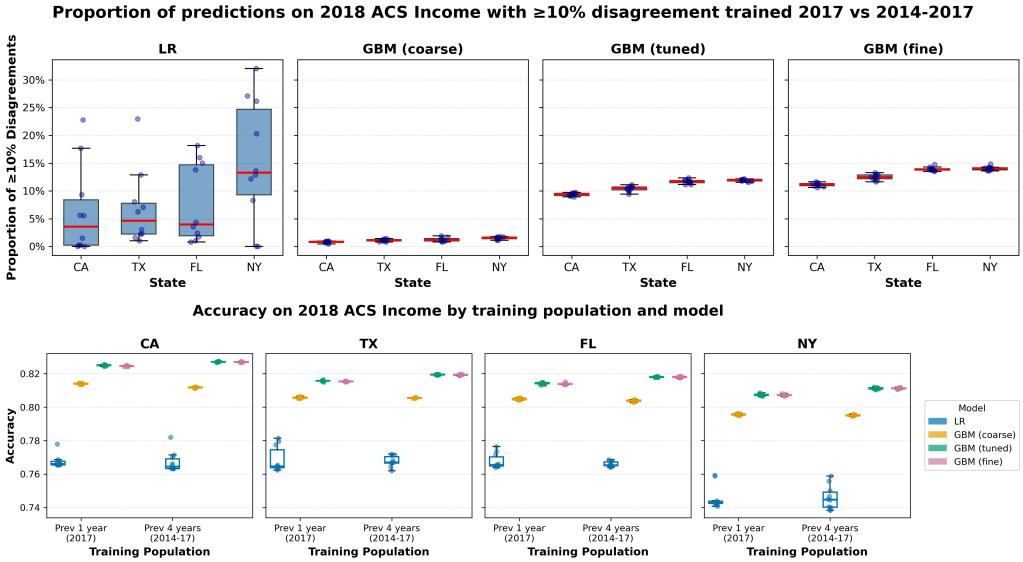

FIGURE 5.1: Top: Model multiplicity based on choice of training population in `ACS Income`, for different model classes on the four largest states. Each dot corresponds to a pair of models, one trained on 2017 data and one on 2014-2017 data (both using the same subset of the 2017 data). Coarse GBM models use default hyperparameters while fine models use those considered optimal on the whole US data in previous work (Cruz and Hardt, 2024; Höltgen and Oliver, 2025), see Appendix 5.7.2.1. The tuned version seems optimal on the four states; here, around 10% of data points in 2018 have at least 10% disagreement in the prediction, depending on whether the models were trained only on the last year or on the last four years. Coarser GBMs show less multiplicity, coarser GBMs show more; LRs results are quite volatile. Bottom: Tuned GBMs significantly outperform GBMs with default settings as well as LR models and have similar performance to the finer GBMs.

will change dramatically when one or two more datapoints are observed. Take the mentioned credit example. Assume that we have access to a look-up table (or oracle) that tells us all the conditional probabilities. Is it sensible to deny people with attributes x_1 a loan because, despite their favourable profile, the default ratio among three datapoints is 2/3? Given that both the data seen in training and the data seen at deployment are thought to be sampled from the empirical distribution, the considered population must cover both the past and the future.[5] So some of the three points may be in the past, and some may be in the future. Therefore, we cannot even argue that predicting the true empir-

use and the metrics we care about; as Nelson Goodman (1972, p. 18) already suspected, 'rather than similarity providing any guidelines for inductive practice, inductive practice may provide the basis for some canons of similarity'.

5 We might want to say that they come from different distributions between which a distribution shift occurs—but this distribution shift can be almost arbitrary if each conditional probability depends only on two or three datapoints.

ical distribution helps us to optimise our decisions in the future: The Gen-D framework fails to account for the phenomenon that datapoints seen in the past will not be seen again in the future: in practice, we sample *without* replacement (for the relevance of this, see also footnote 8 below). A central problem with the conventional framework is, then, that it takes the data to originate from distributions; actually, it is the other way around.

5.3 GEN-DS ARE NOT NEEDED TO UNDERSTAND LEARNING

Now in spite of these theoretical considerations, the Gen-D framework seems to provide a useful framework for modelling and understanding learning. In the following, we demonstrate that the assumption of data-generating distributions is not needed for this. Empirically, we do see that ML models actually do learn patterns, even on human behaviour—so if we do not learn some empirical distribution, then must there not be an abstract true distribution after all? This conclusion is particularly intuitive because the Gen-D framework seems to be the only framework at our disposal to model such stability, via the law of large numbers. We now demonstrate that it is an unwarranted conclusion, both by drawing a historical parallel and by introducing an alternative framework.

With only a little stretching, we can trace precursors to this kind of argument back to the early statisticians in the 19th century like Adolphe Quetelet who, observing, *e.g.*, a 'frightening regularity with which the same crimes are reproduced', tried to find the laws that govern society (Porter, 1986). This also led to a widespread belief in true probabilities which should govern human behaviour according to general laws (Hacking, 1990), very similar to the true data-generating distributions referred to in ML. To explain the stable rates of crimes within countries, Quetelet came up with an abstract construct called the average man ('l'homme moyen'), which, 'defined in terms of the average of all human attributes in a given country, could be treated as the "type" of the nation, the representative of a society in social science comparable to the center of gravity in physics' (Porter, 1986). The notion of a true conditional distribution in ML today is different basically only in the level of granularity: we now have more data such that we can exploit apparent stabilities on a finer level. For then as for now, however, we can find alternative frameworks which capture such stabilities.

One idea is that if we cut a large enough sequence of datapoints with binary labels ('crime', 'no crime') in half (two consecutive years), then for most partitions, the label ratios in both halves will be roughly similar. This can also be framed as drawing from an urn without replacement, where we calculate the

number of possible draws of N out of 2N balls for which the ratio of 'crime' balls is roughly the same as for the remaining N. We capture this through the counting measure (μ in the results below), quantifying the number of admissible draws without any notion of probability. *If* we assigned the same probability to each possible draw (a notion of exchangeability), we could interpret this measure as a probability—but this is an additional assumption, a special case, in this framework. We can extend this to a model of learning, only slightly adapting the work of Vapnik. The following result is similar to Theorem A.2 in Vapnik (1982, p. 170); the difference is that we bound the two notions of generalisation error in finite populations rather than the error when generalising to a true distribution. The proof relies on a modified Symmetrisation Lemma, of which we prove two versions; we chose a slightly simplified version of Vapnik's notation here to facilitate comparisons between results and proofs.

Lemma 5.1 (Symmetrisation Lemma, version 1)**.**
Take an urn with $l + k$ *balls, with* $k \geqslant l > 2/\epsilon$*. Let* $u(\alpha)$ *denote the error rate of* α *on the whole urn. Then*

$$\mu\left[\sup_{\alpha} |u(\alpha) - \nu_{tr}(\alpha)| > \epsilon\right] < 2 \cdot \mu\left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \left(\frac{1}{2} + \frac{l}{k}\right)\epsilon\right]. \tag{5.1}$$

Note that sampling without replacement converges to sampling with replacement for $k \to \infty$. Accordingly, $k \to \infty$ implies $\left(\frac{1}{2} + \frac{l}{k}\right)\epsilon \to \epsilon/2$ which gives Vapnik's original Symmetrisation lemma ('The Basic Lemma', p.168).

Lemma 5.2 (Symmetrisation Lemma, version 2)**.**
Take an urn with $l + k$ *balls, with* $k \geqslant l > 2/\epsilon$*. Let* $u'(\alpha)$ *denote the error rate of* α *on the remaining urn. Then*

$$\mu\left[\sup_{\alpha} \left|u'(\alpha) - \nu_{tr}(\alpha)\right| > \epsilon\right] < 2 \cdot \mu\left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \epsilon/2\right]. \tag{5.2}$$

For $k = l$ we have $u' = \nu_{te}$ and this version is weaker than the previous version's bound with $\epsilon' := \epsilon/2$. given that $2 \cdot |u - \nu_{tr}| = |\nu_{te} - \nu_{tr}|$, which then gives the RHS bound $3/4\epsilon$. This is because above we take the cumulative from $u - \epsilon'/2 = u' - \epsilon/2 - \epsilon/4$, whereas here we take it from $u' - \epsilon/2$. Note that for $k \to \infty$, we get $u' \to u$ and, thus, Vapnik's original Symmetrisation lemma. With these results under our belt, we can now prove the generalisation bound.

Theorem 5.3.
Take a set[6] *of* $l + k$ *datapoints in* $\mathcal{X} \times \{0, 1\}$ *and draw a training set of size* l *without replacement. Consider a hypothesis class* Λ *of finite VC dimension* h*. Denote by* $u(\alpha)$*,*

6 It would be more correct to speak of multi-sets, but we conform with common lingo here.

$\nu_{tr}(\alpha)$, *and* $u'(\alpha)$ *the error rate of* $f_\alpha \in \Lambda$ *on the whole set of* $k+l$ *points, on the training set of* l *points, and on the remaining* k *points, respectively.*

Then for $k \geqslant l > 2/\epsilon$ *with* $\epsilon > 0$*, we can bound the proportion of draws in which the generalisation error exceeds* ϵ *by*

$$\mu\left[\sup_{\alpha\in\Lambda} |u(\alpha) - \nu_{tr}(\alpha)| > \epsilon\right] \;<\; 9\frac{(2l)^h}{h!} \cdot \exp\left(-\epsilon^2 l \left(\frac{1}{2} + \frac{l}{k}\right)^2\right) \tag{5.3}$$

and

$$\mu\left[\sup_{\alpha\in\Lambda} \left|u'(\alpha) - \nu_{tr}(\alpha)\right| > \epsilon\right] \;<\; 9\frac{(2l)^h}{h!} \cdot \exp\left(-\epsilon^2 l/4\right), \tag{5.4}$$

where μ *is the counting measure on possible draws of size* l*.*

At first glance, the successful loss minimisation of ML models across many domains seems to strengthen the case for the Gen-D framework. But as the above result shows, we can understand learning also without invoking a data-generating distribution—indeed, the ideas of pattern recognition and generalisation do not rely on such a distribution at all. The theorem shows that for large enough sample sizes, the error on the sample will be similar to the error on the whole set for the vast majority of possible samples. Note that this provides a more concrete notion of generalisation, featuring data that can (in principle) be seen in the future rather than metaphysical notions of expectations under theoretical distributions; indeed, this relates directly to the generalisation error between training and test set that is commonly investigated in practice. At this point, the reader may ask, all models are wrong, so what? There is some alternative framework that was in some sense already implicit in the Gen-D framework, so what? After all, science commonly works with idealisations (Potochnik, 2017), acting as if the models were true (Appiah, 2017). In the remainder of this paper, we argue that there are good reasons to stop working with thinking in terms of data-generating distributions, at least in social settings.

5.4 GEN-DS FACILITATE MISINTERPRETATIONS

Drawing on both philosophical and theoretical considerations, we have argued that generative distributions are fictions which are not even that useful. In the remainder, we argue that their invocation can even have negative effects in practice, as they can suggest an understanding of ML as an approximation of true probabilities, depending less strongly on particular choices of data collection and training.

5.4.1 *Gen-Ds obscure what is being optimised*

The ML literature often suggests implicitly or explicitly that the goal of learning is to recover the data-generating distribution or to approximate the Bayes-optimal classifier, which is a function of the true conditional $P(Y|X)$. This is indeed appealing if one believes that there is such a true distribution to recover, estimate, or approximate. But, as we have discussed, there is actually no such distribution. As put by a physicist, 'the phrase "estimating a probability" is just as much a logical incongruity as "assigning a frequency"' (Jaynes, 1985, p. 120). The goal of learning, as also formalised in empirical risk minimisation and statistical decision theory, is to predict well on average, to minimise a loss that reflects our preferences and needs (on a population of interest, see footnote 4). Recall the example from Section 5.2 concerning the prediction of credit default, where two out of three people with some x in a population default: There is no true probability that we want to estimate; rather, we want to learn patterns or decision rules that work well on average (see also Höltgen (2026)).

This is not always appreciated nowadays, arguably due to the optimism associated with modern Deep Learning and Big Data. This is reflected in the extremely popular textbook by central figures in this area, which states that '[i]deally, we would like to match the true data-generating distribution p_{data}, but we have no direct access to this distribution' (Goodfellow, Bengio and Courville, 2016, p. 130). In a recent real-world example, the Austrian public employment service rebuked criticism of their logistic regression model, developed to help allocate job training programmes, claimed that its prediction 'follows the real chances on the labour market as far as possible' (as cited by Allhutter et al. (2020)). Such statements, which appear to become more common also with the more recent optimism based on the ever-increasing availability of data and compute (Boyd and Crawford, 2012), mistake error minimisation for an approximation of true probabilities. This has arguably also to do with the use of more complex model classes, as one can theoretically show that in the limit of infinite data, an idealisation related to Gen-D, models like (infinitely big) neural networks or random forests can approximate or will converge to any data-generating distribution.

The role or meaning of probabilities is also relevant to the way we evaluate and select models in practice. As observed by Cynthia Dwork, 'without an answer to this definitional question, we don't even know what it is that the ideal algorithm should satisfy' (Dwork, 2022). The arguments against the Gen-D assumption, in particular, highlight the task dependence of ML-based predictors. In the conventional framework, the conditional distribution $P(Y|X)$ provides the

best predictor for every task: It would 'suffice' to approximate this $P(Y|X)$ and use it for every task that requires the prediction of Y, as this would minimise any proper loss. Indeed, such arguments are often made. However, if there is no true underlying distribution, it becomes difficult to separate the trained predictor from the specific learning tasks. Indeed, much recent work has focused on getting multi-calibrated (Hébert-Johnson et al., 2018) or omni-predictors (Gopalan, Kim and Reingold, 2023), deriving guarantees based on the assumption that we can draw ever more data from the same stable distribution with replacement. The underlying intuition of a 'march towards truth' (attributed to Cynthia Dwork (Roth and Tolbert, 2025)) is clearly in conflict with a central insight of the present paper, that there is no such thing as a true data-generating distribution. It also surfaces in the assumption that multi-calibration–type algorithms like RECONCILE (Roth, Tolbert and Weinstein, 2023) are thought to work 'independent of how rich the feature space is', even up to individual people (Roth and Tolbert, 2025)—which clearly cannot guarantee ever-improving predictions about people with unique profiles that were not represented in the training dataset. The maths checks out, however, because the Gen-D framework assumes that we sample people *with* replacement—which is at odds with assuming that observed events and predicted events have no overlap. We contend that the Gen-D framework is inadequate for capturing the limitations of such approaches.[7]

In practice, the observation that there is no correct predictor that is adequate for every task goes beyond the difficulty of different loss functions leading to different approximations (*i.e.* models): now, there is no true probability that could serve as a mutual reference point for the different models. It also goes beyond the empirical observation of the 'Rashomon effect' or 'predictive/model multiplicity', of which Leo Breimann already said in 2001 that 'its effect on conclusions drawn from models needs serious attention' (Breiman, 2001, p. 206). It is, however, only recently that the community is increasingly grappling with this phenomenon, that there are often many models which perform equally well on average for a given task while disagreeing significantly on individual predictions (Black, Raghavan and Barocas, 2022; Höltgen and Oliver, 2025; Marx, Calmon and Ustun, 2020). Illustrations of this phenomenon on datasets for

7 Recent work has highlighted that multi-calibration is also possible with hardly any assumption, by coupling it with defensive forecasting strategies for the online learning setting (Dwork et al., 2025; Perdomo, 2025). However, the calibration guarantees in this setting only guarantee predicting the base rate (per pre-defined group) while (swap) regret guarantees only serve as a comparison against very specific alternatives. They are all very different to the 'march towards truth' that we criticise here and do not have the discriminative power of forward-looking ML algorithms. Clearly, without making some assumptions about the future, we cannot say how good our predictions will be in absolute terms.

credit default, rearrests, and medical diagnoses in human populations were recently given by (Watson-Daniels, Parkes and Ustun, 2023); they show that discrepancies in individual predictions of $20-40$ percentage points are common for different models with near-optimal loss—with the *same* model class, *same* abstraction, and *same* loss function. In the Gen-D framework, this multiplicity is only an approximation problem, a practical nuisance due to a constrained model class or a small sample size; without this framework, we can see that predictive multiplicity is actually a fundamental conceptual issue (see also Heljakka et al. (2022)).

This is also important for the notion of fairness-accuracy trade-offs (Corbett-Davies et al., 2017; Hardt, Price and Srebro, 2016; Menon and Williamson, 2018). These analyses are based on the Bayes-optimal classifier, that is, they assume a data-generating distribution. Given this framing, fairness considerations are necessarily seen as deviations from optimality. A Empirical Work, however, shows that it is often possible to get less discriminative models without reductions in performance metrics like average accuracy (Black et al., 2024a; Cooper et al., 2024; Höltgen and Oliver, 2025). While the trade-off result could be interpreted as speaking purely about theoretical concepts, they are often thought to also apply to actual models in practice (otherwise, what would be their relevance?): Even fairness researchers sometimes explicitly assert that such results should hold in practice, stating e.g. that considered models 'are likely close to Bayes optimal under the squared loss' (Cruz and Hardt, 2024, p. 2) (echoing Ding et al. (2021)). As the trained models can rarely be considered close to the *empirically* optimal model (otherwise, we would not need machine learning with its inductive biases), such discussions do implicitly or explicitly assume the existence of the data-generating distributions and, thus, of a correct model that is being approximated. Indeed, recent theoretical results that take into account the larger difference between trained models and empirically optimal ones show that there is typically much room for less discriminative but equally accurate models (Höltgen and Oliver, 2025).

Another interesting aspect relating to model performance is that belief in true probabilities may make us overestimate the benefits of individualistic interventions as compared to structural ones. A recent extensive empirical investigation into a programme aimed at preventing students from dropping out of high school in Wisconsin using ML finds that the best way forward would not lie 'in identifying students at risk of dropping out within specific schools, but rather in overcoming structural differences across different school districts' (Perdomo et al., 2025). The former approach *does* seem more powerful, though—if we be-

lieve that students have a true individual risk that is out there for us to find (see also (Perdomo, 2024; Shirali, Abebe and Hardt, 2024)).

5.4.2 *Gen-Ds take abstractions for granted*

So far, we have been discussing the assumption of a true generative distribution for a given input and label space, in particular with the assumption of true conditional probabilities $P(Y|X = x)$. In contrast to that, we now focus on what ML predictions are actually used for—that is, the individual events that we represent through abstractions $x \in \mathcal{X}$. This could, for example, be the probability that you in particular, rather than anyone with your attributes x, will commit a crime within the next year. This relates to the so-called 'reference class problem' for frequentism (Reichenbach, 1949): if we take the probability of some event A to consist in the ratio (or relative frequency) of how often events similar to A do occur, then this will depend on the deployed notion of similarity. An illustrative example is the risk of developing a disease in the next year, say lung cancer: the risk is usually taken to be the ratio of people 'like you' getting the disease. But this ratio will depend on whether we take into account age, gender, country of residence, whether you smoke, what you eat, et cetera. If we take into account too many attributes, we will end up with a single datapoint, you. This again shows that there is no single correct probability, that probabilities are inherently model-dependent.[8]

In this sense, ML models by themselves never provide probabilities for individual events; they always predict probabilities $P(Y = 1|X = x)$ that are conditioned on a certain abstraction. In practice, we automatically map individual events to their representation in the input space $\mathcal{X}$, so this nuance is easily overlooked – especially as literature on ML typically takes the input space $\mathcal{X}$ as given. Consequently, the construction of abstractions, that is, the choice of the input space, is often not discussed enough in ML.[9] However, the choice of how to represent the events we want to predict determines how we carve up the world and, thus, which events are seen as similar. And since learning from experience or data always depends on the notion of similarity (via inductive bi-

8 In an example showing how the Gen-D framework can obscure this, recent work claims to make progress on a version of the reference class problem, yet assuming that we can sample (*with* replacement) from a distribution that already defines (possibly trivial) true probabilities, thereby also *fixing* a level of abstraction (Roth and Tolbert, 2025; Roth, Tolbert and Weinstein, 2023). As a more general point, the freedom to choose an adequate level of abstraction need not be seen as a 'problem' to solve (Höltgen, 2026).

9 This arguably has to do with the observation that 'everyone wants to do the model work, not the data work' (Sambasivan et al., 2021).

ases favouring smooth predictors in the case of ML), this, in turn, influences the resulting predictions. This can be seen as a generalisation of the reference class problem to cases where we do not just predict the average label of all events that we represent through the same abstraction: The outputs of more complicated prediction methods that are based on more flexible notions of similarity also depend on the mode of abstraction—which allows the modeller to chose a suitable abstraction for a given task (Höltgen, 2026).

We have argued previously that it is already problematic to even assume true conditional probabilities $P(Y|X = x)$ for modelling. It seems even more problematic to argue that some selection of attributes x should be the 'correct' representation of John Doe for predicting, *e.g.*, whether he will commit a crime or find a job within the next year. And even with a fixed abstraction, the predicted probabilities depend on other choices about the data, such as the considered time frame, et cetera. So, again, there is no true probability to be found—but in contrast to the former, there can be a lot of disagreement on how to model such more complicated events and how to calculate probabilities.

Given that it is quite clear that the choice of abstraction matters for the predictions, it is quite remarkable how little this choice is sometimes discussed (depending on who you talk to). As a result, decision makers sometimes choose to include demographic attributes like gender in practice, but we still do not have a solid understanding of when this leads to more discrimination; and a lack of established best practices necessarily leads to a lack of accountability. Recent work (Höltgen and Oliver, 2025) suggests that more such attributes do not necessarily lead to better predictions, but may exacerbate discrimination (see our Figure 5.2)—contrary to the myth of Big Data (Boyd and Crawford, 2012), but also contrary to our idealised theoretical models. We, thus, argue that the framing of ML problems in terms of a fixed space $\mathcal{X} \times \mathcal{Y}$ with a true joint distribution has contributed to this: Partly because it is outside this problem framing, and partly because the true distribution already conveys that whatever features one selects, one can approximate true probabilities. It has been noted that because of the way ML is done, 'the abstraction is taken as given and is rarely if ever interrogated for validity. This is despite the fact that abstraction choices often occur implicitly, as accidents of opportunity and access to data' (Selbst et al., 2019, p. 60). The Gen-D framework, thus, induces us to underestimate the inferential leap in identifying a conditional probability $P(Y|x)$ to commit a crime for your attributes x with 'your' probability to commit a crime: your data was not sampled from a distribution, it was constructed as a profile of you.

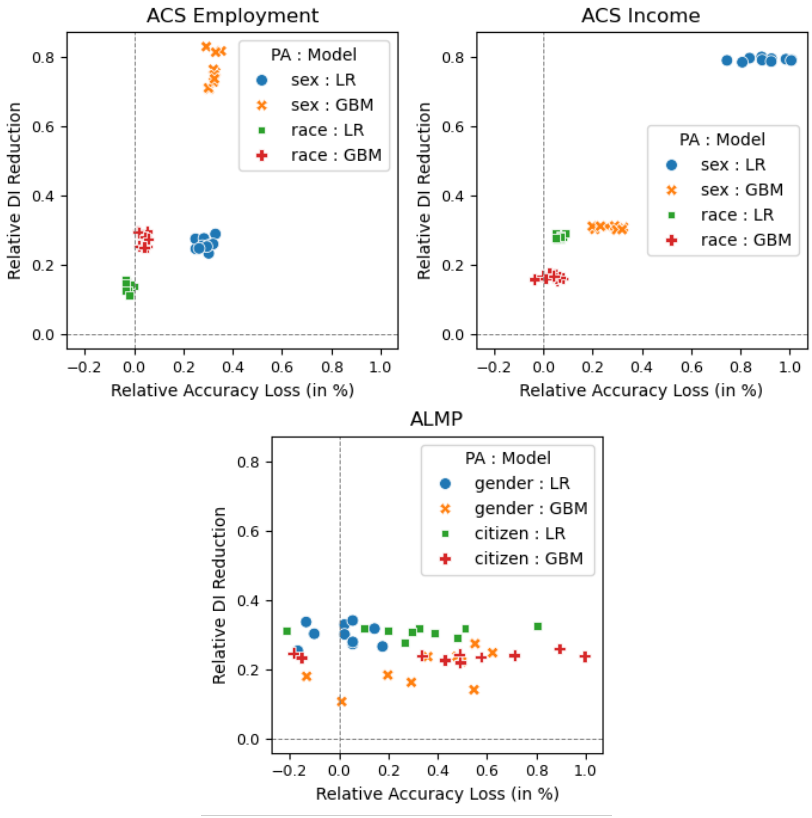

FIGURE 5.2: Reproduced from Höltgen and Oliver (2025) with the authors' permission: Relative reduction in Accuracy vs in Disparate Impact (DI) across three large datasets, for logistic regression (LR) and gradient boosted trees (GBM). Each dot corresponds to a model pair with the same train/test split, one with and one without the protected attribute (PA). Omitting the PA reduces DI by 15-80% (y-axis) but hardly reduces—and sometimes improves—the accuracy (x-axis). The original publication (Höltgen and Oliver, 2025) contains additional details.

5.5 TRUE PROBABILITIES AND THE ILLUSION OF OBJECTIVITY

Algorithmic predictions are often presented as objective; in the previously mentioned example on job training allocations, predictions were said to aim for 'the *highest objectivity possible* in the sense that the predicted integration likelihood [...] follows the real chances on the labour market as far as possible' (as cited by Allhutter et al. (2020), emphasis added). This perception of objectivity is partly an effect of quantification in general (Porter, 1995), but partly also specific to ML applications, where the data are often thought to 'speak for themselves' (Boyd and Crawford, 2012; Moss, 2021).

If there were true probabilities, then getting a low(er) average loss could be seen as an indication that one's predictions get close(r) to these true probabilities, vindicating the model irrespective of the choices that went into designing

and training it. The choices would only serve the purpose of getting close to the true probabilities, and with enough data, we may think that we get close enough to make the specific choices less relevant. If, however, there is no such true probability to approximate, no such indirect vindication is possible; hence, the choices need to be justified more explicitly (Höltgen and Oliver, 2025). Rather than seeing models as constructed on top of available data, even critical work often repeats the phrasing (from a data science textbook, in this case) that statistical models are 'estimating' individual values of interest such as 'a crime location or offender' (Strikwerda, 2021, p. 423) or 'an individual's level of risk' (Jacobs and Wallach, 2021, p. 382). Crucially, the idea that machine learning models approximate individual probabilities contributes to understating the relevance of specific modelling choices; as documented by Moss (2021), this is 'how applied machine learning researchers contribute to the idea that numbers can speak for themselves' (p. 21). Based on this, 'the products of machine learning come to instead appear as separate from their labor [...], not only automatic or magic, but also natural and inevitable' (Moss, 2021, p. 13). Indeed, this idea is directly reflected in the predominant theoretical understanding of machine learning. This idea of approximating objective values is particularly dangerous for probabilistic predictions, as—in contrast to predicting observable attributes like height—we do not observe individual prediction errors. By highlighting that probabilities in general are constructed rather than discovered, we contribute to previous work which argues that algorithmic predictions are not the objective arbiters of truth that they are sometimes taken to be.

Indeed, many 'value-laden choices' (Allhutter et al., 2020) by ML engineers go into designing such systems without later being discussed as such.[10] Because the creators of such models often do not discuss their choices, letting the results 'speak for themselves', 'the products of machine learning come to instead appear as separate from their labor – [...] not only automatic or magic, but also natural and inevitable' (Moss, 2021, p. 13). This concerns not only the construction of models but also of data—as argued in (Leonelli, 2019), 'there is no such thing as "raw data", since data are forged and processed through instruments, formats, algorithms, and settings that embody specific theoretical perspectives on the world'. The deployment of Machine Learning systems for consequential decisions, thus, needs actual justification in relation to the data and modelling choices. This means that these choices need to be more openly discussed and recorded—recent years have seen suggestions for standardised documentation

10 This fits well into the older literature on values in science more generally (Douglas, 2000; Rudner, 1953). Also the choice of 'ontology' (*i.e.* categories, quantities to measure, ...) in science depends on the specific goals (Danks, 2015) and non-epistemic values (Ludwig, 2016).

(which are common in other areas of engineering) to record such choices as well as intended uses for models (Mitchell et al., 2019) and datasets (Gebru et al., 2021; Hutchinson et al., 2021).

It should also be noted that minimising average loss is, first and foremost, meant to be useful for the decision-maker and the population as a whole; it is harder to justify it from the individuals' perspective if it is not possible to claim that their 'objective risk' is being estimated. For example, the choice of abstraction is often more difficult to justify from the individual's perspective, as it determines with whom you are considered equivalent (and who is considered similar, via inductive biases). We are, however, not arguing here categorically against making decisions about people based on data that was collected from other people. After all, that is also something that human decision-making relies on. What we want to caution against is a premature justification of decisions informed by ML-based systems based on the perceived objectivity of such systems, which is reinforced by a naïve view on probability. We are not samples from distributions, we are the real thing.

5.6 CONCLUSION

In this paper, we have problematised the notion of true data-generating distributions as it often appears in the context of Machine Learning. We emphasised that probabilities in general are always constructed rather than discovered (Höltgen, 2026). We have demonstrated that in the case of ML, the ubiquitous assumption of a true data-generating distribution (Gen-D) is not always a benign one, especially when people are involved. In particular, it contributes to the air of objectivity that often circulates in predictions based on data science. As an alternative, we suggested a more parsimonious framework for understanding learning in social settings based on finite populations. In addition to the epistemic modesty it conveys, it can also lead to new intuitions and insights. While working with probability distributions is sometimes considered more elegant, this not only comes at the price of clouding the modelling assumptions[11] but is also less expressive (Meng, 2022). In contrast, modelling finite populations directly requires empirically testable modelling assumptions, which can pave the way for more accountability in model development.

Based on this work, we would like to see further research into three directions in particular. One direction relates to the choice of abstraction/representation,

11 It has also been shown that in practice, statements made in the language of probability more easily mislead people than the same statements made in terms of relative frequencies in populations (Gigerenzer, 2008, Chapter 9).

that is, of the input space/covariates, as highlighted in Section 5.4.2. Although it is quite clear that such choices have a large impact on the resulting predictions, it seems that—for reasons we discussed—very little research so far has addressed this issue. We, thus, encourage researchers as well as practitioners to follow Lily Hu's 'call to arms' (Hu, 2023), to think more about what she calls 'variable construction' and we call the choice of abstraction. Another direction is to further develop theory within the finite population framework; this will sometimes involve easy adaptations, as in Theorem 5.3, but open up more possibilities, e.g., by going beyond i.i.d. sampling assumptions via non-probability sampling as developed in the survey sampling literature (Meng, 2022).

The third direction concerns the limits of our models: While the standard story suggests that there are distributions waiting for us everywhere, not all tasks are created equal in practice. We, thus, need more empirical insights into which domains and tasks and quantities are actually stable to admit fruitful predictions. We will need to check the stabilities that the systems offer against the requirements of our prediction models and tasks (footnote 4), to decide which models (if any) are suitable in which setting. Many systems work remarkably well without us understanding why, but they also often break down unexpectedly, especially in the social world, sometimes with serious consequences for affected populations (Wang et al., 2024). Establishing best practices for stating concrete and domain-specific inductive assumptions can be one element in avoiding more such failures, but also pave the way for more accountability. Assuming a stable distribution by default, and rhetoric like 'approximating the current state' obscures the need for this.

The choice of framework is, in general, more relevant for model evaluation than model training: The evaluation depends directly on the framing of the problem, especially on how performance on current data is thought to inform future performance. A departure from the Gen-D framing should not be understood as a departure from validation, but as a call to think more concretely about the relationships between training, validation, and deployment populations, rather than assuming they are sampled from the same distribution (or from distributions within an arbitrary ϵ-ball on some abstract divergence measure). We highlighted that the idea of 'recovering the true data-generating distribution', is, in most cases, misguided—the aim is, rather, to find a model that performs well for some choice of loss function and aggregation scheme. The goal of ML should be to find a good model, not the true model; this seems to have been clearer in earlier days when Machine Learning was more often understood as Pattern Recognition. Note also that the aggregation scheme for losses need not—and, arguably, often should not—be the average, even though

it is easy to work with and explicitly used in expected risk minimisation (and also in this paper). In many settings, we are interested in other distributional properties that pay more attention to worst-case performance, such as the class of 'fairness risk measures' (Williamson and Menon, 2019).

The choice of training algorithm, on the other hand, appears to be more independent of the chosen framework and has largely eluded rigorous theoretical justification. Deep Learning is not really motivated by any particular formal framework—indeed, seminal papers introducing e.g. Transformers (Vaswani et al., 2017) make no reference to data-generating distributions. For the most part, training then depends on the problem formulation only indirectly, via the evaluation criterion: we choose training methods that work well for the chosen evaluation criterion. Still, looking at the topic from a different angle can be a fresh breeze that provides insights into why and when such models work well. We would like to stress again that the purpose of the paper is not to advocate for a full-blown rejection of the Gen-D framework in general, as it seems adequate in less complex settings like gambling.

Nancy Cartwright argues that 'extending our favourite theories to treat new problems of new kinds' is not dangerous only 'when there is good empirical evidence for the promise of the proposed approach' and 'not objectionable so long as we are clear about what we are doing and what it is going to cost us, and we are able to pay the price' (Cartwright, 1999, p. 333). We have argued that the costs of pretending that there are data-generating distributions in social settings are high and that there are no real benefits (beyond convention/inertia). When it comes to decision-making about people, there is too much risk of contributing to claims of objectivity and the evasion of a thorough justification of modelling choices. The language we use and assumptions we make around ML systems are particularly relevant for the literature on fair and just deployment of algorithms; all too often, societal concerns are seen as departures from simply approximating the true probabilities. We take issue with 'simply', 'approximating', and 'the true probabilities'.

5.7 APPENDIX

5.7.1 *Proofs*

We consider a population of $k+l$ datapoints of which l are taken for the training set. $\nu_{tr}(\alpha)$ and $u(\alpha)$ are the error ratios of model α on the train set and the whole urn, respectively. We consider a theoretical test set with l samples and error rate $\nu_{te}(\alpha)$. We omit '(α)' when the model is clear or irrelevant.

We define u' to be the error rate in the remaining urn after taking the train set, *i.e.*

$$u' = \frac{l+k}{k}u - \frac{l}{k}v_{tr} = u + \frac{l}{k}(u - v_{tr}) \tag{5.5}$$

since $u(k+l) = v_{tr}l + u'k$.

For the proof of the theorem, we need modified, finite-urn versions of Vapnik's Symmetrisation Lemma.

Lemma 5.1 (Symmetrisation Lemma, version 1)**.**
Take an urn with $l+k$ balls, with $k \geqslant l > 2/\epsilon$. Let $u(\alpha)$ denote the error rate of α on the whole urn. Then

$$\mu\left[\sup_{\alpha} |u(\alpha) - v_{tr}(\alpha)| > \epsilon\right] < 2 \cdot \mu\left[\sup_{\alpha} |v_{te}(\alpha) - v_{tr}(\alpha)| > \left(\frac{1}{2} + \frac{l}{k}\right)\epsilon\right]. \tag{5.6}$$

Proof. Drawing $2l$ balls without replacement and partitioning them into train and test set is the same as first taking l for the train set without replacement and then taking l for the test set since we are only concerned with numbers of combinations.

The proof is based on Vapnik's but improves on a naïve application by taking into account that the test set is sampled from a changed urn with a slightly different error ratio.

We give a lower bound for the RHS by showing that draws that break the LHS bound have at least a $1/2$-probability that they also break the RHS bound. Put differently, of all combinations that break the LHS bound, at least $1/2$ also break the RHS bound.

Consider a draw for which there is an α^* with $u(\alpha^*) - v_{tr}(\alpha^*) > \epsilon$ (analogous for $v_{tr}(\alpha^*) - u(\alpha^*) > \epsilon$).

So assuming that $u - v_{tr} > \epsilon$ we see that $u' - v_{te} < \epsilon/2$ would guarantee

$$\begin{aligned} & v_{te} - v_{tr} && (5.7)\\ & = (v_{te} - u') + (u' - u) + (u - v_{tr}) && (5.8)\\ & > -\epsilon/2 + (u' - u) + \epsilon && (5.9)\\ & = \epsilon/2 + (u' - u) && (5.10) \end{aligned}$$

and thus, since

$$u' - u = \left(u + \frac{l}{k}(u - v_{tr})\right) - u = \frac{l}{k}(u - v_{tr}) > \frac{l}{k}\epsilon \tag{5.11}$$

we get

$$|v_{te} - v_{tr}| > \left(\frac{1}{2} + \frac{l}{k}\right)\epsilon. \tag{5.12}$$

This means that

$$\mu(u' - \nu_{te} < \epsilon/2 | u - \nu_{tr} > \epsilon) \cdot \mu \left[\sup_{\alpha} |u(\alpha) - \nu_{tr}(\alpha)| > \epsilon \right]$$
$$< \mu \left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \left(\frac{1}{2} + \frac{l}{k} \right) \epsilon \right]. \tag{5.13}$$

Now $\mu(u' - \nu_{te} < \epsilon/2 | u - \nu_{tr} > \epsilon)$ is the probability that for an urn with k balls and $ku' > ku + l\epsilon$ red ones, we draw l balls of which at least $(u' - \epsilon/2)l$ are red. Now

$$\mu(u' - \nu_{te} < \epsilon/2 | u - \nu_{tr} > \epsilon) > 1/2, \tag{5.14}$$

given that the expected number of balls drawn is $u'l$ and $\epsilon l/2 > 1$ guarantees that there is an integer between $u'l$ and $(u' + \epsilon/2)l$. Now (5.13) and (5.14) taken together prove the lemma. □

Lemma 5.2 (Symmetrisation Lemma, version 2)**.**
Take an urn with $l + k$ balls, with $k \geqslant l > 2/\epsilon$. Let $u'(\alpha)$ denote the error rate of α on the remaining urn. Then

$$\mu \left[\sup_{\alpha} |u'(\alpha) - \nu_{tr}(\alpha)| > \epsilon \right] < 2 \cdot \mu \left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \epsilon/2 \right]. \tag{5.15}$$

Proof. Same proof as in Vapnik' Symmetrisation Lemma but with $u'(\alpha)$ (see (5.5)) instead of $P(\alpha)$:

We show that of all draws where some α breaks the LHS bound, at least half also break the RHS bound.

Consider draws that break the LHS and fix an α^* for which the LHS bound is broken, and assume w.l.o.g. $u' - \nu_{tr} > \epsilon$ (the reverse case is analogous).

Then if $u' - \nu_{te} < \epsilon/2$, we get

$$\nu_{te} - \nu_{tr} = (\nu_{te} - u') - (u' - \nu_{tr}) > \epsilon - \epsilon/2 = \epsilon/2. \tag{5.16}$$

As the reverse case is analogous, we can bound

$$\mu(u' - \nu_{te} < \epsilon/2 | u' - \nu_{tr} > \epsilon) \cdot \mu \left[\sup_{\alpha} |u'(\alpha) - \nu_{tr}(\alpha)| > \epsilon \right] \tag{5.17}$$
$$< \mu \left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \epsilon/2 \right]. \tag{5.18}$$

Now $\mu(u' - \nu_{te} < \epsilon/2 | u - \nu_{tr} > \epsilon)$ is the probability that for an urn with k balls and $ku' > ku + l\epsilon$ red ones, we draw l balls of which at least $(u' - \epsilon/2)l$ are red. Now

$$\mu(u' - \nu_{te} < \epsilon/2 | u - \nu_{tr} > \epsilon) > 1/2, \tag{5.19}$$

given that the expected number of balls drawn is $u'l$ and $\epsilon l/2 > 1$ guarantees that there is an integer between $u'l$ and $(u' + \epsilon/2)l$.

Now (5.18) and (5.19) taken together prove the lemma. □

Theorem 5.3.
Take a set[12] *of* $l + k$ *datapoints in* $\mathcal{X} \times \{0, 1\}$ *and draw a training set of size* l *without replacement. Consider a hypothesis class* Λ *of finite VC dimension* h*. Denote by* $u(\alpha)$*,* $\nu_{tr}(\alpha)$*, and* $u'(\alpha)$ *the error rate of* $f_\alpha \in \Lambda$ *on the whole set of* $k + l$ *points, on the training set of* l *points, and on the remaining* k *points, respectively.*

Then for $k \geqslant l > 2/\epsilon$ *with* $\epsilon > 0$*, we can bound the proportion of draws in which the generalisation error exceeds* ϵ *by*

$$\mu\left[\sup_{\alpha\in\Lambda} |u(\alpha) - \nu_{tr}(\alpha)| > \epsilon\right] < 9\frac{(2l)^h}{h!} \cdot \exp\left(-\epsilon^2 l \left(\frac{1}{2} + \frac{l}{k}\right)^2\right) \tag{5.20}$$

and

$$\mu\left[\sup_{\alpha\in\Lambda} \left|u'(\alpha) - \nu_{tr}(\alpha)\right| > \epsilon\right] < 9\frac{(2l)^h}{h!} \cdot \exp\left(-\epsilon^2 l/4\right), \tag{5.21}$$

where μ *is the counting measure on possible draws of size* l*.*

Proof. It is sufficient to bound

$$\mu\left[\sup_{\alpha} |\nu_{te}(\alpha) - \nu_{tr}(\alpha)| > \epsilon\right] < m^S(2l) \cdot 3\exp\left(-\epsilon^2 l\right) \tag{5.22}$$

for two sets of size l the same way as in Vapnik's proof of Theorem A.2 and use the respective Symmetrisation Lemma. The growth function $m^s(l)$ is bounded by $m^S(l) < 1.5\frac{l^h}{h!}$, see Theorem 6.6 in (Vapnik, 1982, p. 154). □

Note that both versions converge to Vapnik's result for $k \to \infty$, as this means $l/k \to 0$ and $u' \to u$, respectively. Also note that our second version is stronger (the bound is tighter) than the original version in the sense that the LHS is larger while the RHS is the same.

5.7.2 *Experiments*

5.7.2.1 *More details on the experimental setup*

For the results reported in Figures 5.1 an 5.3, we used `scikit-learn` default parameters for *LR* except setting `max-iter=1000`. We used default parameters of `scikit-learn` for *GBM (coarse)*; for *GBM (fine)*, we took the parameters reported as optimal for the whole US data of `ACS Income` by Höltgen and Oliver (2025), similar to those of Cruz and Hardt (2024); we tuned them to try improve performance while making it coarser, resulting in *GBM (tuned)*. Details on the GBM configurations are given in Table 5.1. Each seed uses 50% of the respective

12 It would be more correct to speak of multi-sets, but we conform with common lingo here.

year for training. For the model pairs that are compared in Figure 5.1, all data on which the 2017 model is trained is also used to train its respective 2014-2017 partner model.

TABLE 5.1: Hyperparameter configurations for the used GBM settings; *GBM (coarse)* is the `scikit-learn` default.

Hyperparameter	GBM (coarse)	GBM (tuned)	GBM (fine)
`n_estimators`	100	400	500
`max_depth`	3	8	10
`max_leaf_nodes`	None	50	50
`min_samples_leaf`	1	500	500

5.7.2.2 *Additional experiments on model degradation*

We ran additional experiments where models were trained on a random subset of the 2014 data of the four biggest states (10 subsets for 10 models) and tested on subsequent years. We observe performance degradation for each subsequent year (Figure 5.3), comparable to that reported on the whole US data by Ding et al. (2021).

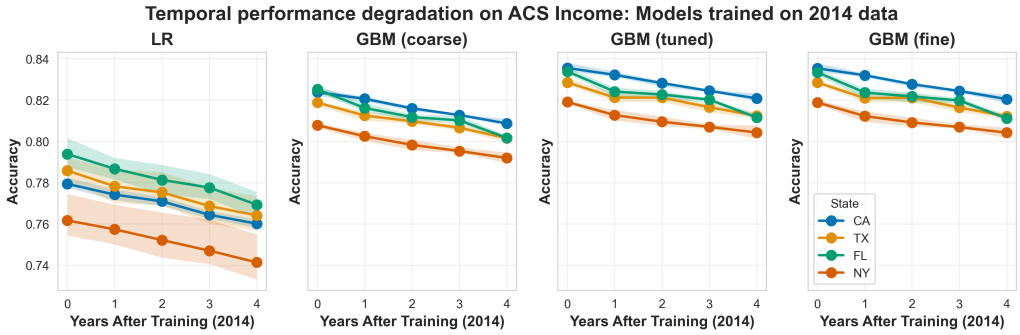

FIGURE 5.3: Model degradation for various models trained on 2014 `ACS Income` data of different states and tested on subsequent years. Coloured shadows visualise min and max over 10 seeds.

ACKNOWLEDGEMENTS

We would like to thank Rabanus Derr, Nina Effenberger, and Christian Fröhlich for helpful feedback on earlier versions.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number

2064/1—Project number 390727645 as well as the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Benedikt Höltgen.

6

RECONSIDERING FAIRNESS THROUGH UNAWARENESS FROM THE PERSPECTIVE OF MODEL MULTIPLICITY

Normally screws are so cheap and small and simple you think of them as unimportant. But now, as your Quality awareness becomes stronger, you realize that this one, individual, particular screw is neither cheap nor small nor unimportant.

— Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance*

AUTHOR CONTRIBUTIONS

Benedikt Höltgen had the idea for the project based on discussions with Nuria Oliver. The project, including conceptualisation, theoretical results, and empirical analysis, was conducted by Benedikt Höltgen under the supervision of Nuria Oliver. The paper was written by Benedikt Höltgen, with valuable feedback from Nuria Oliver.

ABSTRACT

Fairness through Unawareness (FtU) describes the idea that discrimination against demographic groups can be avoided by not considering group membership in the decisions or predictions. This idea has long been criticized in the machine learning literature as not being sufficient to ensure fairness. In addition, the use of additional features is typically thought to increase the accuracy of the predictions for all groups, so that FtU is sometimes thought to be detrimental to all groups. In this paper, we show both theoretically and empirically that FtU can reduce algorithmic discrimination without necessarily reducing accuracy. We connect this insight with the literature on Model Multiplicity, to which we contribute with novel theoretical and empirical results. Furthermore, we illustrate how, in a real-life application, FtU can contribute to the deployment of more equitable policies without losing efficacy. Our findings suggest that FtU is worth considering in practical applications, particularly in high-risk

scenarios, and that the use of protected attributes such as gender in predictive models should be accompanied by a clear and well-founded justification.

6.1 INTRODUCTION

Nurtured by the ever-increasing availability of data and compute, as well as a widespread fondness of quantification in general (Porter, 1995) and Big Data in particular (Boyd and Crawford, 2012), machine learning-based systems are increasingly used to assist decision making in a variety of areas of society, including in high-stakes domains such as healthcare, finance, law enforcement, and the provision of social services. Leveraging data is often seen not only as a way to improve decisions but also as a way to avoid human bias and enhance objectivity and fairness. However, there is mounting evidence of a variety of problems that can arise in different contexts as a result of this trend (Wang et al., 2024). One such problem is unintended algorithmic discrimination, which can arise for different reasons, including biased training data, flawed model assumptions, unequal representation of different demographic groups, or misuse of the models. As a result, algorithmic decisions may systematically disadvantage certain individuals or communities grouped by their protected attributes, such as gender, race, age, or religion. These concerns have prompted a substantial body of research focused on understanding, detecting, and mitigating algorithmic biases (Barocas, Hardt and Narayanan, 2023). A seemingly straightforward approach to preventing algorithmic discrimination across groups is the exclusion of protected attributes from the set of features that the model can "see". This idea is known as *Fairness through Unawareness* or FtU (Barocas, Hardt and Narayanan, 2023). However, the algorithmic fairness literature has shown that avoiding the use of protected attributes in the models is generally not sufficient to prevent discrimination (Calders and Verwer, 2010; Pedreshi, Ruggieri and Turini, 2008): other variables can encode group membership information and thus their use could lead to discrimination. Furthermore, FtU is often criticized not only for being an unreliable approach to achieve fairness, but also for potentially reducing the predictive performance of the model for all groups (Corbett-Davies et al., 2023).[1]

Such results align with a common intuition in the Big Data era: that more data (and attributes) lead to better, fairer, and more objective decisions (Boyd and Crawford, 2012). A recent practical example of this way of thinking can be found in Austria, where gender was used as an input in an algorithm designed

1 It has been noted that such a reduction in accuracy need not lead to a reduction in utility (Coots et al., 2023).

to allocate job training programs to unemployed individuals. Officials dismissed criticisms regarding potential discriminatory outcomes by arguing that the algorithm's predictions reflect the actual chances individuals face in the labour market (as cited in Allhutter et al. (2020)). More generally, many arguments in favour of including protected attributes rely on the idea that machine learning models aim to estimate "true" probabilities from data, which is a questionable assumption (Höltgen and Williamson, 2026b). As a result of this assumption, modelling choices may be accepted with limited scrutiny or justification, despite their significant implications (Moss, 2021). However, there is growing evidence that policy outcomes are highly sensitive to modelling decisions, including the choice of thresholds, input variables, and specific design parameters (Bach et al., 2023; Kern et al., 2024). In parallel, emerging research on *model multiplicity* highlights that different models trained on the same data can perform similarly in terms of accuracy, yet differ substantially in their behaviour, particularly with respect to fairness (Black, Raghavan and Barocas, 2022). This perspective highlights that there is no clear hierarchy between models in terms of accuracy and that modelling choices need to be checked against alternative choices.

In light of these concerns, we revisit the commonly held scepticism toward *Fairness through Unawareness* (FtU). While we do not suggest that FtU offers a simple solution to achieve algorithmic fairness, we demonstrate that, contrary to prevailing assumptions, omitting protected attributes does not necessarily harm predictive performance and can contribute to more equitable outcomes. To substantiate these claims, we analyse the relationship between data, model class, and fairness, both from a theoretical and from an empirical perspective. While we derive general insights, we place particular emphasis on logistic regression because of its broad and extensive use in practice to build models from tabular data. Regarding fairness, we focus on disparate impact (DI) or statistical/demographic parity, which compares the positive classification rates for different groups according to their protected attribute. Our findings suggest that including protected attributes can be especially problematic in the context of logistic regression, as their use by such models can easily exacerbate disparities between protected groups. By linking the concepts of FtU and model multiplicity, we argue for a shift in perspective: fairness should not be seen as a constraint or a trade-off, but as a deliberate modelling choice that can lead to more equitable outcomes without necessarily sacrificing predictive performance.

In sum, the main **contributions** of this work are as follows:

- We analyse the relationship between disparate impact and base rate difference in the data both for any classifier and for logistic regression. In the

case of logistic regression, we demonstrate the risk of exacerbated unfairness depending on the classification thresholds.

- We contribute to the understanding of model multiplicity and algorithmic fairness by providing mathematical results that characterise the potential for multiplicity and connect it to fairness: (1) general upper bounds on model multiplicity and changes in disparate impact which are tight for unconstrained models; and (2) lower bounds on achievable reductions in disparate impact based on FtU for logistic regression.
- We empirically demonstrate how unaware models can yield similar levels of accuracy than aware models with significant improvements in fairness in multiple settings.
- We highlight the practical implications on our work by discussing a real-world example related to employment programs in a European country.

The **structure** of this article is as follows. In Section 6.2, we provide an overview of related work on fairness, less discriminatory alternatives, and model multiplicity. In Section 6.3, we highlight the importance of modelling choices by demonstrating that there is no straightforward connection between disparate impact and differences in base rates, even assuming well-calibrated models. We also show that disparate impact can often be higher than the base rate difference for *aware* models (*i.e.* models using protected attributes), especially for models based on logistic regression. In Section 6.4, we frame our work from the perspective of model multiplicity. We provide new theoretical upper bounds and relate them to disparate impact and to fairness through unawareness. In the case of logistic regression, we also provide theoretical lower bounds for reducing disparate impact. We illustrate our insights empirically in Section 6.5, showing that FtU can provide fairer models with almost equal accuracy across multiple datasets, protected attributes, and model classes. Section 6.6 illustrates the impact of our work in a real-world scenario where an aware logistic regression model was used to assign unemployed individuals to job training programs. We show how in this case an unaware model would be equally performative but fairer. Finally, the main conclusions and lines of future work are described in Section 6.7.

6.2 RELATED WORK

6.2.1 *Algorithmic fairness and disparate impact*

Under the influence of legal, economic, and philosophical notions of discrimination as well as several high-profile investigations into existing software, an extensive body of work on algorithmic fairness has introduced a variety of metrics to quantify and address discrimination in machine learning systems (Barocas, Hardt and Narayanan, 2023). One line of work focuses on error rate disparities, *i.e.*, the differences in false positive or false negative rates across demographic groups. However, some scholars have argued that error rates should serve as *diagnostic tools* rather than direct targets for fairness interventions (Barocas, Hardt and Narayanan, 2023). Another influential line of work draws on legal and policy traditions, particularly anti-discrimination law in the United States, introducing the concept of *disparate impact*—also referred to in the machine learning literature as *demographic or statistical parity* (Barocas and Selbst, 2016; Feldman et al., 2015). Disparate Impact focuses on algorithmic outcomes rather than errors: it compares the rates at which individuals from different demographic groups receive positive classifications (*e.g.*, job offers, university admissions, loan approvals). If one group consistently receives favourable outcomes at a lower rate than other groups, this may be evidence of systemic bias.[2] While disparate impact has the advantage of aligning closely with legal standards and public perceptions, it also faces criticism, particularly in machine learning contexts. Enforcing strict parity in classification outcomes may lead to a reduction in accuracy and thus overall "utility" (Hardt, Price and Srebro, 2016), especially when base rates differ across groups. This trade-off can result in decisions that, while statistically balanced, may harm individuals or reduce efficiency without delivering clear benefits (Corbett-Davies et al., 2023). Moreover, socio-technical critiques point out that measures such as disparate impact that focus on classification rates, just like error-focused measures, only consider model properties and may fail to result in just outcomes when actually deployed (Kuppler et al., 2022; Selbst et al., 2019). Nonetheless, disparate impact has a more straightforward connection to the real-world consequences of models than error rates. In this work, we assume that reducing disparate impact is often desirable as long as it does not imply a notable reduction in accuracy, broadly in line with calls for non-ideal approaches to fairness (Fazelpour and Lipton, 2020) and comparative

2 These outcome disparities are sometimes quantified by means of the *disparate impact ratio* and evaluated using heuristics like the "80% rule"—which, however, is only a "crude test" (Raghavan and Kim, 2024).

(rather than transcendental) approaches to justice (Sen, 2010). Indeed, this perspective reflects the original motivation behind the notion of disparate impact; in contrast, its criticism in algorithmic fairness often concerns the full elimination of disparate impact, which reflects its common treatment as a mathematical formula that is exchangeable with other metrics.

6.2.2 *Less discriminatory alternatives and model multiplicity*

In legal doctrine under U.S. anti-discrimination law, the concept of disparate impact is closely related to the idea of a *Less Discriminatory Alternative*: a policy or decision-making process that leads to disparate impact can be justifiable as long as there exists no alternative that meets the same legitimate objectives while producing less discriminatory outcomes. This principle plays a central role in Title VII of the Civil Rights Act and has informed a wide range of employment and civil rights cases (Barocas and Selbst, 2016). In the context of machine learning, less discriminatory alternatives can be operationalized as a fairness-aware criterion for model selection: if two models perform similarly regarding predictive accuracy but one model has lower disparate impact than the other, the fairer model could be considered as the less discriminatory alternative and thus would be both ethically preferable and potentially legally required in regulated environments (Barocas and Selbst, 2016).

Although central to disparate impact, the concept of less discriminatory alternatives has only recently gained attention in the ML literature, particularly in the context of *model multiplicity* (MM) (Black et al., 2024b; Gillis, Meursault and Ustun, 2024). MM describes the phenomenon where multiple different models can achieve comparable accuracy on the same task. While the underlying idea—sometimes called *Rashomon effect*—has been known for decades (Breiman, 2001), only in recent years has it become a focus of systematic investigation (Marx, Calmon and Ustun, 2020). Despite promising recent work (Semenova et al., 2023), this area of research is still in its early stages.

An intriguing aspect of MM is that models with similar accuracy can exhibit very different behaviours regarding robustness, fairness, or interpretability (Black, Raghavan and Barocas, 2022; Coston, Rambachan and Chouldechova, 2021; Rudin et al., 2024). By virtue of MM, it is often possible to select fairer models without losing accuracy, *i.e.* less discriminatory *algorithms* (LDAs) (Black et al., 2024a). In such cases, the existence of accurate and fairer alternatives underscores the need to examine the modelling pipeline beyond aggregate performance metrics. From a fairness perspective, MM implies that selecting one model over another is not only a neutral performance-based decision, as it can have a

differential real-world impact across demographic groups. In this paper, we investigate how Fairness through Unawareness (FtU) can, in some cases, provide a less discriminatory alternative without requiring explicit fairness optimisation, which has been investigated in other works (Coston, Rambachan and Chouldechova, 2021; Gillis, Meursault and Ustun, 2024). We examine whether models that exclude protected attributes from their input can still maintain competitive predictive performance while reducing disparate impact, particularly in the context of logistic regression, a widely used model in real-world scenarios.

6.3 CALIBRATED AWARE MODELS CAN EXACERBATE INEQUALITY

> *"Predictive models trained with supervised learning methods are often good at calibration [...] But calibration also means that by default, we should expect our models to faithfully reflect disparities found in the input data."*
> (Barocas, Hardt and Narayanan, 2023, p. 21)

Reflected in this quote from the most prominent textbook on algorithmic fairness is the frequent assumption that disparities present in the data will be carried over into calibrated machine learning models by default. While this assumption may hold true for probabilistic models, it does not apply to classifiers derived from these models. In this section, we demonstrate that, perhaps counterintuitively, decisions based on calibrated probabilities can actually worsen the inequalities reflected in the base rates of labels. We also show that this effect is especially pronounced in policies relying on logistic regression models.

We assume an input space $\mathcal{X}$ and a, possibly empirical, joint probability distribution P over random variables (X, Y) which can take binary values in $\mathcal{X} \times \{0, 1\}$. Take a predictor $f : \mathcal{X} \to [0, 1]$ and a corresponding classifier

$$F : \mathcal{X} \to \{0, 1\},\ x \mapsto \mathbb{1}[f(x) \geqslant 0.5]. \tag{6.1}$$

Assume two groups that form a partition $\{\mathcal{A}, \mathcal{B}\}$ of $\mathcal{X}$.[3] The purpose of this section is to analyse how the disparate impact (DI)

$$\mathsf{DI}(F) := \mathbb{E}_P[F(X) \mid\ X \in \mathcal{A}] - \mathbb{E}_P[F(X) \mid\ X \in \mathcal{B}] \tag{6.2}$$

relates to the difference in base rates (DBR)

$$\mathsf{DBR} := \mathbb{E}_P[Y \mid\ X \in \mathcal{A}] - \mathbb{E}_P[Y \mid\ X \in \mathcal{B}]. \tag{6.3}$$

In this paper, we assume that $\mathcal{B}$ is the disadvantaged group, *i.e.*, $\mathsf{DBR} \geqslant 0$.

3 This is the case when the input features contain a protected attribute encoding group membership, i.e., for group-aware models.

6.3.1 *General case*

The first observation is that for calibrated predictors, the label rates in the data are reflected in the means of the predictors. In the standard sense of Chouldechova (2017), a predictor f is *calibrated* on both groups if

$$\forall v \in [0,1], * \in \{\mathcal{A}, \mathcal{B}\}: \quad \mathbb{E}_P[Y \mid f(X) = v, X \in *] = v. \tag{6.4}$$

Now, if f is calibrated in this sense, then the average prediction of f on either group trivially equals the average outcome for that group (simply by integrating over v):

$$\forall * \in \{\mathcal{A}, \mathcal{B}\}: \quad \mathbb{E}_P[f(X) \mid X \in *] = \mathbb{E}_P[Y \mid X \in *]. \tag{6.5}$$

While all predictors that are calibrated on a given dataset will have the same mean, equal to the base rate, these different predictors can lead to different classification rates. We illustrate this with a simple toy example.

Example 6.1. *Take* $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ *with* P *defined via* $\forall x \in \mathcal{X} : P(x) = 0.25$ *and* $P(Y = 1 \mid X = x) = 0.4, 0.4, 0.55, 0.7$ *for* $x = x_1, x_2, x_3, x_4$*, respectively. Then the predictors* f_1, f_2, f_3 *as defined in Table 6.1a are all calibrated in the standard sense (without groups) while leading to very different classifications (Table 6.1b).*

TABLE 6.1: Left (table a): Calibrated predictors for Example 6.1, where $p(x) = P(Y = 1 \mid X = x)$ is the true probability. f_1 is given by joining x_3 with x_1 and x_2; f_2 given by joining x_1 with x_3 and joining x_2 with x_4; f_3 is given by the constant base rate predictor. All of these predictors are calibrated. Right (table b): Classifiers derived by the models in table a, where B is the Bayes optimal classifier.

a	x_1	x_2	x_3	x_4
$p(x)$	0.4	0.4	0.55	0.7
$f_1(x)$	0.45	0.45	0.45	0.7
$f_2(x)$	0.475	0.55	0.475	0.55
$f_3(x)$	0.5125	0.5125	0.5125	0.5125

b	x_1	x_2	x_3	x_4
$B(x)$	0	0	1	1
$F_1(x)$	0	0	0	1
$F_2(x)$	0	1	0	1
$F_3(x)$	1	1	1	1

We turn next to the quantities of interest: disparate impact and difference in base rates. First, we consider the most general case, with arbitrary f, α, and P. It is important to note that disparate impact depends on the ratio of predictions above/below the decision threshold, whereas the base rate difference depends on the means of the predictions (assuming calibration). We can formalise this in terms of the distributions of the predictions for each group, $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$, where

$$P_*(v) := P(f(X) = v \mid X \in *) \quad \text{for} \quad v \in [0,1],\ * \in \{\mathcal{A}, \mathcal{B}\}. \tag{6.6}$$

Then, for random variables $V_a, V_b \sim P_{\mathcal{A}}, P_{\mathcal{B}}$ describing the group-wise predictions, the disparate impact is given by

$$\begin{aligned} DI(F) &= P(F(X) = 1 \mid X \in \mathcal{A}) - P(F(X) = 1 \mid X \in \mathcal{B}) && (6.7)\\ &= P_{\mathcal{A}}(\{V_a \geqslant 0.5\}) - P_{\mathcal{B}}(\{V_b \geqslant 0.5\}) && (6.8) \end{aligned}$$

and the difference in base rates by

$$DBR = \mathbb{E}_{P_{\mathcal{A}}}[V_a] - \mathbb{E}_{P_{\mathcal{B}}}[V_b]. \tag{6.9}$$

The difference in base rates can intuitively be understood as the balance of how much and how far the probability mass is moved to the left vs the right of the decision boundary when moving from $P_{\mathcal{B}}$ to $P_{\mathcal{A}}$.[4] The disparate impact, on the other hand, only depends on the balance of how much probability mass passes over the decision threshold.

The base rate difference and the disparate impact can be directly related under specific conditions, for example, when $P_{\mathcal{A}}$ is higher on one side of the threshold and $P_{\mathcal{B}}$ is higher on the other (proofs can be found in Appendix 6.8.2):

Proposition 6.2 (Relation between Base Rates and Disparate Impact)**.** *Assume* f *is calibrated as in (6.4), satisfying* $\forall v \in [0, 0.5) : P_{\mathcal{B}}(v) \geqslant P_{\mathcal{A}}(v)$ *and* $\forall v \in [0.5, 1] : P_{\mathcal{A}}(v) \geqslant P_{\mathcal{B}}(v)$*. Then for* F *defined as in (6.1),* $DI(F) \geqslant DBR$*.*

6.3.2 *Special case: Logistic Regression*

We now investigate the relationship between disparate impact and the difference in base rates in the case of logistic regression. Logistic regression is a widely used type of statistical method for modelling the probability of a binary outcome based on one or more predictor variables. Its popularity is due to its simplicity, interpretability, and effectiveness in tackling classification problems. The model class consists of linear models with a sigmoid output σ,

$$f : \mathbb{R}^n \to [0, 1],\ (x_1, ..., x_n) \mapsto \sigma\left(c_0 + \sum_i c_i x_i\right) \tag{6.10}$$

with $\mathcal{X} \subset \mathbb{R}^n$ and the sigmoid function defined by $\sigma : z \mapsto \frac{1}{1+e^{-z}}$.

Training a logistic regression model consists of finding the values of the model's coefficients $c_0, ..., c_n \in \mathbb{R}$ that best fit the data by minimising a logistic loss function (equivalent to maximising the likelihood) using optimisation techniques like gradient descent. If a protected attribute is part of the input space,

4 This means that if $P_{\mathcal{A}}$ 'dominates' $P_{\mathcal{B}}$ in the sense of $\exists T : [0, 1] \to [0, 1]$ with $T(v) \geqslant v\ \forall v \in [0, 1]$ s.t. $T(V_b) \sim P_{\mathcal{A}}$, then the base rate difference equals the earth mover (Wasserstein-1) distance.

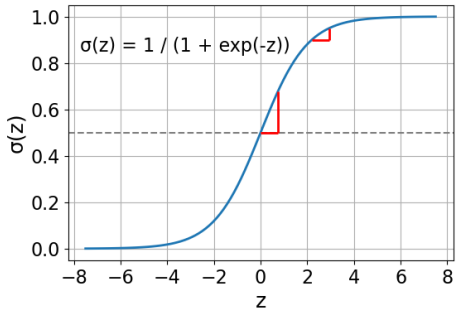

FIGURE 6.1: Plot of the sigmoid function used in logistic regression. Note that the slope is steepest around $z = 0$ and satisfies $\sigma(0) = 0.5$.

there will be a corresponding coefficient $c_i \in \mathbb{R}$ which determines how the protected attribute affects the (pre-sigmoid) logit function, $\hat{f}(x) := c_0 + \sum_i c_i x_i$. Therefore, for binary protected attributes where x_i can be 0 or 1, being part of the disadvantaged group ($x_G = 1$) implies adding a term c_G to the logit without further interaction with the rest of the variables.

We focus on the case where $P(Y \mid X, G = *)$ differs significantly between groups, as this makes the existence of a fixed global coefficient c_G particularly problematic.[5] A difference in base rates means that, if the model is calibrated for both groups (6.5), the coefficient of the variable that corresponds to the group attribute will penalise all members of the disadvantaged group because the coefficient for that group will reduce the logit by a fixed value c_G, before applying the sigmoid, which can then contribute to an even larger disparate impact. To understand why, consider the shape of the sigmoid function displayed in Figure 6.1: The curve is steeper around the classification threshold at $\sigma(z) = 0.5$ than for other values of z, as highlighted by the red lines. Therefore, small changes in z near the classification threshold would lead to significant changes in the value of the sigmoid.

Consider the following example with a logistic regression classifier: Two individuals have feature vectors x^b and x^a that are identical except for their group membership, such that person b belongs to group $\mathcal{B}$ and person a belongs to group $\mathcal{A}$. Therefore, the only element that differs between the two individuals' predictions is the coefficient corresponding to the variable that denotes which group they belong to. Changes in this coefficient will shift the logit value z. Imagine the output of the logistic regression function (predicted probabilities) for each of them is $\sigma(z^b) = 0.95$ and $\sigma(z^a) = 0.9$. To obtain the logit values z^a and z^b, we need to apply the inverse sigmoid $z = \sigma^{-1}(p) = \ln(\frac{p}{1-p})$. In the example, $\sigma^{-1}(0.95) \approx 2.94$ and $\sigma^{-1}(0.90) \approx 2.20$, such that their difference is ≈ 0.75. This

5 In the case where $P(Y \mid X, G = *)$ is the same across groups but $P(X \mid G = *)$ is different, the choice of model class may not have as immediate an effect. However, this choice still affects the resulting predictive distribution, which has an effect on the disparate impact as shown above.

difference comes from the assumed difference in base rates between groups in this example. This suggests that group $\mathcal{A}$ has a group offset that raises the logit by 0.75 when compared to group $\mathcal{B}$ for the same input features. Now, let's consider an individual in group $\mathcal{B}$ that has a predicted probability right at the classification threshold at $\sigma(z) = 0.5$ (therefore, logit $z = 0$); and another individual with exactly the same feature vector but belonging to group $\mathcal{A}$. The logit would increase by 0.75 because of the group effect, such that $z_a = z_b + 0.75 = 0 + 0.75$, $\sigma(0.75) \approx 0.68$. This means that the same features lead to higher predictions for members of $\mathcal{A}$ due to the group offset: for every individual in group $\mathcal{A}$ with scores between 0.5 and 0.68, there is a counterpart in $\mathcal{B}$ who would fall below the threshold, even though their other features are identical.

Thus, a global group coefficient that is meant to account for overall base rate differences between groups can lead to large local effects, especially in the vicinity of the decision threshold, due to the non-linear shape of the sigmoid function. In the example, even a moderate difference of 0.05 closer to the margins corresponds to a large difference of 0.18 near the threshold, which can lead to a large disparate impact.

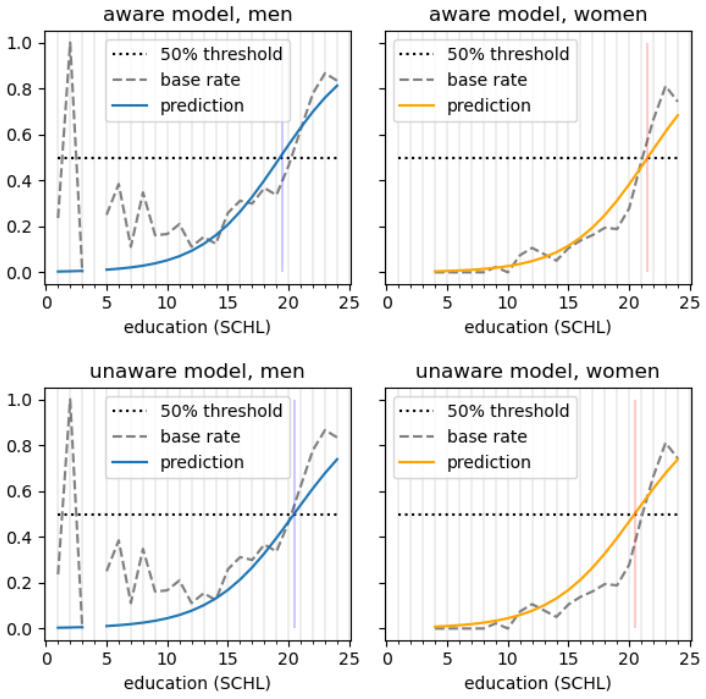

FIGURE 6.2: Predictions for aware and unaware logistic regression models on the ACS Income NY dataset. The aware model uses education and sex as input variables. The unaware model only considers education, measured in years. Colored vertical lines correspond to the 50% cut-off. Note how the aware logistic regression model (top) requires much higher education values for women (yellow/red) than for men (blue) to be selected. This phenomenon is not observed in the unaware model.

This effect can be further illustrated with an example based on real data, namely the US Census data from New York, obtained from the ACS Income dataset from the `folktables` package (Ding et al., 2021) (see Section 6.5.1). In this case, the target variable is whether a person has an annual income of at least 50k. For illustration purposes, we restrict the logistic regression model to include only two variables: the most predictive attribute (education, in number of years) and the protected binary attribute (sex). Two logistic regression models using default hyperparameters are trained: one *aware* model, including education and sex, and one *unaware* model that uses education as the only input variable. The results are depicted in Figure 6.2: The aware model "pushes" the predictions for men and women apart, meaning that men need an education score of 20 (equivalent to an associate's degree), whereas women need an education score of 22 (equivalent to a Master's degree) to obtain a positive classification. Income for men is over-predicted around the threshold (the blue line being above the grey dashed base rate line), whereas for women it is under-predicted (the yellow line being below the grey dashed base rate line). Conversely, in the case of the unaware model, which does not include sex as an input variable, the blue line on the left and the orange line on the right are necessarily the same. In this case, men are predicted correctly at the threshold while women are slightly over-predicted (the yellow line being above the grey dashed base rate line). Overall, this results in a similar accuracy for the unaware model (69.9%), compared to the aware model (70.0%), while significantly reducing the disparate impact (more details can be found in Appendix 6.8.1.3).[6]

Before empirically investigating larger datasets, we take a step back and theoretically analyse the effect as an instance of model multiplicity—the phenomenon that multiple models can have similar accuracy while differing in other properties, such as fairness.

6.4 LESS DISCRIMINATORY ALGORITHMS THROUGH UNAWARENESS

In this section, we show that large differences between models with similar accuracy can be explained theoretically and that datapoints with predictions close to the decision threshold are particularly important for this. We provide the first tight upper bounds on model multiplicity and, based on that, derive bounds on the difference in disparate impact. We also provide a lower bound specifically for logistic regression classifiers.

6 While the single feature case is an unrealistic illustration, the fact that the accuracy is not much lower than when using all features indicates that the feature drives most of the prediction. This already suggests that similar results occur in full datasets; we turn to this in Section 6.5.

6.4.1 *General upper bounds on model multiplicity*

We begin by establishing general results on model multiplicity and then explore their implications for disparate impact and unaware models that do not use the protected attribute. Let P denote the empirical distribution over X, Y, describing the available data; based on this we define the conditional $p(x) : x \mapsto P(Y = 1 \mid X = x)$ and marginal measure $\mu(U) : 2^{\mathcal{X}} \to [0,1], U \mapsto \int_U P(x)dx$. Furthermore, let $B : \mathcal{X} \to \{0,1\}$ be a Bayes-optimal classifier (with $B(x) = 0$ for $p(x) < 0.5$ and $B(x) = 1$ for $p(x) > 0.5$). Now in line with the closest previous work (Black, Raghavan and Barocas, 2022), we define L as the 0-1-loss

$$L(y_1, y_2) := \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{otherwise.} \end{cases} \tag{6.11}$$

We define the accuracy of a model F as

$$Acc(F) := 1 - \mathbb{E}_P[L(F(X), Y)] \tag{6.12}$$

and the disagreement between two models F, G as

$$d(F, G) := \mu(\{x \in X : F(x) \neq G(x)\}). \tag{6.13}$$

For $\mathcal{F}_{\mathcal{X}} := 2^{\mathcal{X}}$ the set of binary classifiers on $\mathcal{X}$, let

$$R_\epsilon(F) := \{G \in \mathcal{F}_{\mathcal{X}} : Acc(G) \geqslant Acc(F) - \epsilon\} \tag{6.14}$$

be the ϵ-Rashomon set of F. For $\lambda \in [0, 0.5]$, let

$$U(\lambda) := \{x \in \mathcal{X} : |p(x) - 0.5| = \lambda\} \quad \text{and} \quad m(\lambda) := \mu(U(\lambda)) \tag{6.15}$$

measure how much each distance from 0.5 occurs, which is non-zero only on the set of occurring distances $\Lambda := \{|p(x) - 0.5| : x \in \mathcal{X}\}$. Then we define new quantities

$$e(\lambda) := \sum_{\lambda' < \lambda} 2\lambda' \cdot m(\lambda') \quad \text{and} \quad \lambda_\epsilon := \arg\max_{\lambda \in \Lambda} \{\lambda : e(\lambda) \leqslant \epsilon\}, \tag{6.16}$$

which are illustrated in Figure 6.7 in Appendix 6.8.2. Intuitively, $e(\lambda)$ is the additional error if for all points with label ratios closer to 50% than λ, the suboptimal classification is taken, and λ_ϵ is the largest λ such that this error does not exceed ϵ. With these definitions at hand, we can prove a tight upper bound on model multiplicity, which is proved through a lemma on randomised classifiers. The dependence of our tight bound on λ_ϵ and $e(\lambda_\epsilon)$ demonstrates how important input space regions with a near-50% label ratio are for model multiplicity.

Proposition 6.3 (Upper Bound on Multiplicity)**.**

$$\max_{F \in R_\epsilon(B)} d(F, B) \leqslant \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon} + \sum_{\lambda < \lambda_\epsilon} m(\lambda). \tag{6.17}$$

This bound is the tightest possible bound with only access to the cumulative distribution of $p(x)$.

Lemma 6.4. *The bound (6.17) can be achieved in any setting for every* ϵ *by a* randomised *classifier defined as*

$$F_\epsilon(x) = \begin{cases} 1 - B(x) & \text{for } |p(x) - 0.5| < \lambda_\epsilon \\ 1 - B(x) \text{ with probability } \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon \cdot m(\lambda_\epsilon)} & \text{for } |p(x) - 0.5| = \lambda_\epsilon \\ B(x) & \text{for } |p(x) - 0.5| > \lambda_\epsilon. \end{cases} \tag{6.18}$$

This means the maximum is achieved if the predictions with $p(x)$ closest to 0.5 are flipped until the accuracy decreases by ϵ. All proofs, as well as illustrations of these results and the quantity $e(\lambda)$ can be found in Appendix 6.8.2.

We now relate this bound to a **previous result**: Black, Raghavan and Barocas (2022) provide a formal upper bound on how much a model can disagree with another model when the difference in accuracy is controlled. We can also recover their result in the proof of their Theorem A.2, that

$$d(F, B) \leqslant \frac{Acc(B) - Acc(F)}{2c} \tag{6.19}$$

under the assumption[7]

$$|p(x) - 0.5| \geqslant c > 0 \; \forall x \in \mathcal{X}. \tag{6.20}$$

Corollary 6.5. *Under assumption (6.20),*

$$\max_{F \in R_\epsilon(B)} d(F, B) \leqslant \frac{\epsilon}{2c}. \tag{6.21}$$

While the above bounds involved an optimal classifier, we can also give bounds for **two non-optimal models**.

Proposition 6.6 (Multiplicity Bound for Non-optimal Models)**.**
We can bound the disagreement between two classifiers $F \in R_\epsilon(B)$ *and* $G \in R_\delta(B)$ *via*

$$d(F, G) \leqslant \max_{H \in R_{\epsilon+\delta}(B)} d(H, B). \tag{6.22}$$

7 Black et al. assume $|p(x) - 0.5| > c$ but it is not necessary to assume that the inequality is strict.

That is, the maximum possible disagreement between F and G is no greater than the largest disagreement that any model H within accuracy drop $\epsilon + \delta$ could have with B itself. We now connect this bound to **disparate impact**.

Proposition 6.7 (Upper Bound on Disparate Impact)**.** *For two classifiers* F *and* G*, the difference in their disparate impact can be bounded based on their disagreement and the size of the smaller group:*

$$|\mathrm{DI}(F) - \mathrm{DI}(G)| \leqslant \frac{d(F, G)}{\min_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *)}. \tag{6.23}$$

That is, the potential difference in disparate impact between two models increases with their disagreement, and it is inversely proportional to the size of the smallest protected group. We can also specify a bound for the context of **Fairness through Unawareness**. The above bound is not tight when an unaware model is involved, unless for all $\bigcup_{\lambda<\lambda_\epsilon} U(\lambda)$, the two groups are separable in $U(\lambda)$ by the protected attribute alone. If this is not the case, then the unaware model will give the same decision for some $x_1, x_2 \in \bigcup_{\lambda<\lambda_\epsilon} U(\lambda)$ from different groups; so flipping the decisions on both x_1 and x_2, which is necessary for equality in (6.17), will not lead to a proportional rise in DI, *i.e.*, there is no equality in (6.23). However, we can show that half the bound is achievable for some ϵ:

Proposition 6.8 (Achievable DI for Unaware Models)**.** *Fix* P*, an unaware model* F*, and some* $\epsilon > 0$*. Then there is a model* $G \in R_\epsilon(F)$ *with*

$$|\mathrm{DI}(F) - \mathrm{DI}(G)| \geqslant \frac{1}{2 \cdot \max_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *)} \max_{H \in R_{\epsilon+\delta}(B)} d(H, B). \tag{6.24}$$

6.4.2 *A lower bound for logistic regression*

We now turn to the specific case of LR models. The following result provides a lower bound on the reducible DI for a fixed change in accuracy. This bound depends on the coefficient of the protected attribute and the calibration plots for both groups, including bin sizes. Similar to the results presented above, it crucially depends on changes in classification decisions near the decision threshold, highlighting the importance of data points near the decision threshold.

Proposition 6.9 (Lower Bound on DI for LR)**.** *Let* $F : \mathcal{X} \to \{0, 1\}$ *be an aware classifier based on an LR model* f *with coefficient* $c_G < 0$ *for the PA, i.e., the disadvantaged group gets a pre-sigmoid penalty of* c_G*. Further, let*

$$Q(c_G) := \{x \in \mathcal{B} : f(x) \in [\sigma(c_G), 0.5]\} \tag{6.25}$$

denote the set of points that are in Group b and *get predictions between* $\sigma(c_G)$ *and* 0.5 *by* f. *Then the corresponding unaware classifier* F' *based on setting* $c_{PA} \leftarrow 0$ *in* F *has accuracy bounded by*

$$|\mathrm{Acc}(F) - \mathrm{Acc}(F')| \leqslant 2\sigma(-c_G) \cdot P(X \in Q(c_G)) \tag{6.26}$$

and disparate impact given by

$$\mathrm{DI}(F') = \mathrm{DI}(F) - P(X \in Q(c_G))/P(X \in \mathcal{B}) \tag{6.27}$$

if group b *is still disadvantaged in* F' *(i.e.* $\mathrm{DI}(F) \geqslant P(X \in Q(c_G))/P(X \in \mathcal{B})$*), and otherwise*

$$|\mathrm{DI}(F) - \mathrm{DI}(F')| \leqslant P(X \in Q(c_G))/P(X \in \mathcal{B}). \tag{6.28}$$

The focus on one specific unaware model provides a loose but already interesting lower bound on the possible reduction of DI: As a quantitative example, consider the ACS Income dataset discussed in the following section. In this setting, a logistic regression model assigns a coefficient of -0.86 for the protected attribute sex. We observe that the proportion of women is $P(X \in \mathcal{B}) = 0.48$ and $P(X \in Q(c_G)) = 6.9\%$ is the set of women whose predicted probabilities fall within the range $[0.5, 0.70)$, corresponding to the sigmoid of 0.86, i.e., $\sigma(0.86) = 0.70$. If we were to flip decisions in this region, the accuracy would change by at most 0.7 percentage points either down or up, while the disparate impact would decrease by at least 14.5 percentage points.[8]

Note that this result only represents one possible model within the $(2\sigma(-c_G) \cdot P(X \in Q(c_G)))$-Rashomon set, and therefore serves as a lower bound on the achievable DI reduction. It is unlikely to be optimal since it optimises the other coefficients 'under the assumption' that c_G will be used; it also only modifies the predictions for women, leaving the men's unchanged. In summary, we have provided general upper bounds on disparate impact reduction within some ϵ-Rashomon set that are achievable for unconstrained model classes and we have a naïve lower bound for logistic regression. In the next section, we empirically investigate how unaware and aware models compare on real datasets.

6.5 EMPIRICAL FINDINGS

After the theoretical results above, we now empirically analyse the effect of Fairness through Unawareness (FtU) on Disparate Impact (DI) and Accuracy.

8 With a base accuracy of 77,4% and a DI of 0.052, this translates to a relative accuracy change of 0.009 vs a relative reduction in DI of 0.64—see Figure 6.3.

We consider three datasets, with different model classes and protected attributes. All code, is available at `https://github.com/ben-hoeltgen/ftu-mm/`, including for Section 6.3.2 above.

6.5.1 *Setup*

1. Models: We analyse the behaviour of two model classes, logistic regression (LR) and gradient-boosted decision trees (GBM), via the popular `scikit-learn` package (more details in Appendix 6.8.1.2).
2. Data: We perform experiments on three datasets: The ACS Income and ACS Employment datasets from the `folktables` package (Ding et al., 2021) from the US Census and the active labour market policy (ALMP) evaluation dataset from Switzerland (Lechner et al., 2020). Table 6.4 in Appendix 6.8.1.1 contains a summary of the data. In ACS Income and ACS Employment, the target variables are whether a person's income is above 50k and whether they are employed, respectively. We consider two binary protected attributes: sex (male/female) and race (white/black). Note that the US Census asks for sex rather than gender, which is reflected in the derived dataset. Race is self-indicated in the census, and we restrict to the white/black categories.[9] In the ALMP dataset, our target variable is whether a person will be employed for at least 6 months within the next two years. We consider two binary protected attributes: gender (male/female) and citizenship (Swiss/foreign). In line with previous literature (Zezulka and Genin, 2024), we exclude caseworker information and restrict data to German-speaking cantons.

6.5.2 *Results*

The main results, shown in Figure 6.3, indicate that Accuracy remains similar between aware and unaware models (*i.e.*, models that exclude the protected attribute), while DI is often significantly reduced in the unaware models, yielding fairer models with similar levels of accuracy.[10] The results on the ALMP dataset show general multiplicity, whereas the results on the ACS datasets exhibit less multiplicity. This is not surprising, given that the number of data points is much

9 We want to emphasise that race categories are an over-simplification of the complex notion of racialization. Their use should always be contextualised and avoided where possible (Doh et al., 2025).

10 This result is consistent with previous analysis on the German unemployment data, where it was observed that "in- or excluding protected attributes has little effect on overall performance" (Kern et al., 2024). However, this observation is not further discussed there.

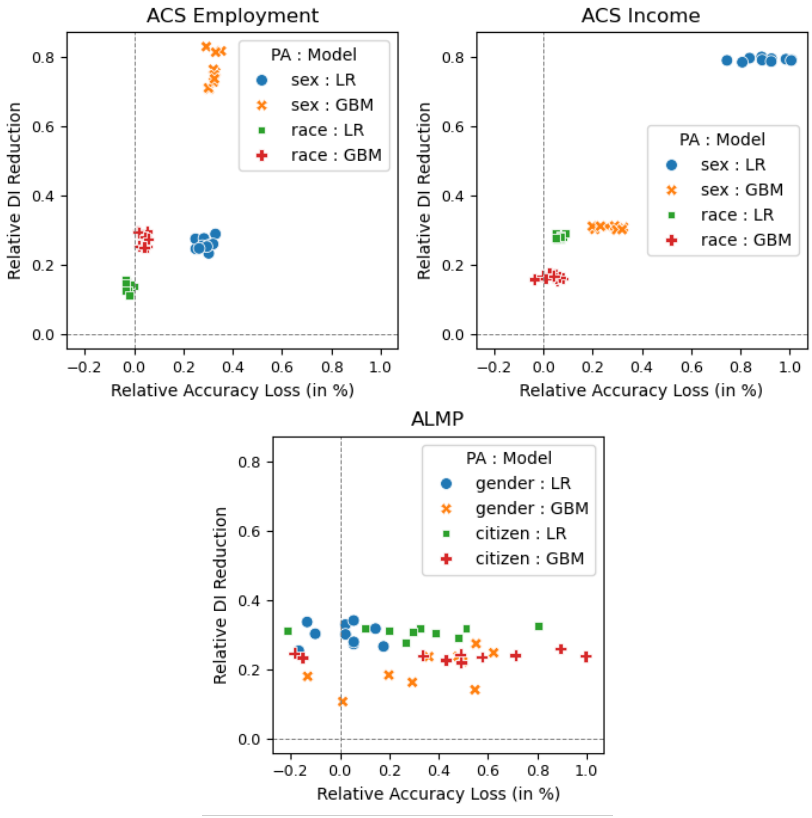

FIGURE 6.3: Relative reduction in Accuracy vs in Disparate Impact (DI) across three datasets, with two model classes (LR: logistic regression and GBM: gradient boosted trees) and two protected attributes (PAs) each. Each dot corresponds to the difference in Accuracy and DI between the aware and the corresponding unaware model trained on the same random train/test split. Note how the unaware model achieves a reduction in accuracy which is always below 1% and sometimes it even increases, while yielding a reduction in DI that is much more substantial (between $15-80\%$).

larger in the ACS datasets than in the ALMP dataset (Table 6.4). The fact that the observed results hold on such large datasets suggests that the effect does not simply disappear with more data.

When considering race as the protected attribute in the ACS datasets, the reduction in disparate impact is generally smaller than with sex, but reaches around 20% across all settings at virtually no loss in accuracy. For settings where sex is the protected attribute, there is an interesting difference in the results on the ACS Income vs the ACS Employment datasets. On ACS Income, the largest reduction in DI ($\approx$ 80%) is obtained with the unaware LR model (consistent with, but slightly exceeding, the lower bound of Section 6.3.2, see footnote 8). In contrast, on ACS Employment, the largest reduction of DI (also $\approx$ 80%) corresponds to the unaware GBM. This result can be explained by two effects that are observable in Figure 6.4. First, aware LR can lead to particularly high dispar-

ate impact, as discussed in Section 6.3.2. This effect is particularly pronounced when sex is the protected attribute. For instance, in the case of the ACS Income dataset, the LR model leads to a disparate impact that far exceeds the base rate difference in the data (indicated by the red line). Second, in the ACS Employment dataset, the LR model with sex as the protected attribute actually results in higher positive classification rates for women than for men, suggesting that women may have stronger predictive features for employment than men. As a result, the change in disparity between men and women is more noticeable with LR than with GBM (see Figure 6.4), even though the absolute DI values change less (see Figure 6.3). This difference between ACS Employment and ACS Income can be attributed to the lower base rate differences in the former (see Table 6.4).

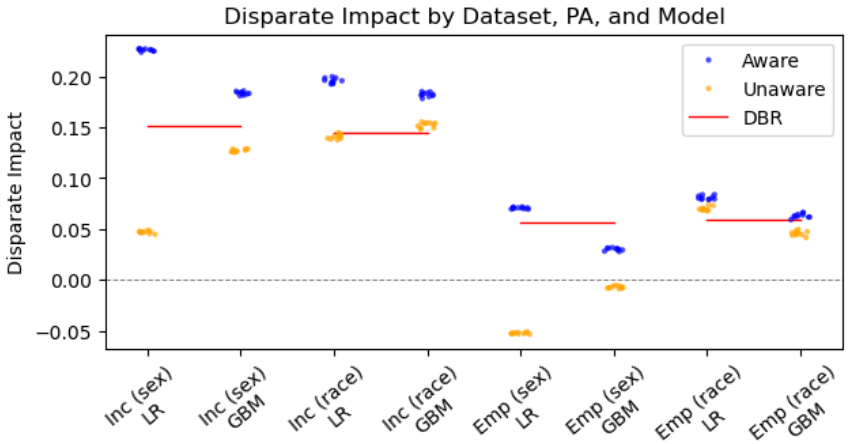

FIGURE 6.4: Disparate impact for the aware and unaware models in different settings. The red lines indicate the difference in base rates in the data. Note how the unaware models achieve lower disparate impact than the aware models.

6.6 DISCUSSION: REAL-WORLD USE CASE

We have shown that including protected attributes like sex or gender in models can lead to substantially less equitable classifications, often without meaningful gains in accuracy—and sometimes without improving accuracy at all. However, what ultimately matters is not just how models classify individuals, but the real-world impact that these systems have when used to guide decisions. In other words, what matters are the *policies* that these models inform (Kuppler et al., 2022). As others have pointed out, *"[i]t is policies and their effects that are just or unjust; 'fair' predictors can both support unjust policies and undermine just policy"* (Zezulka and Genin, 2024, p. 1984). Machine learning models are increasingly

used in resource allocation, where limited goods or opportunities—like counselling or training programs—are distributed based on risk predictions, often through threshold-based policies.[11] An important question is whether predictions are even helpful in such contexts (Perdomo, 2024; Perdomo et al., 2025; Shirali, Abebe and Hardt, 2024). Furthermore, there is the concern that predicting risk irrespective of treatment may not suffice since this risk does not necessarily correlate with treatment effectiveness (Sharma and Wilder, 2025; Zezulka and Genin, 2024). With these caveats, we now turn to examine the impact of Fairness through Unawareness in a real-world application that has attracted considerable attention in Europe in recent years: the statistical profiling of job seekers, which is increasingly used in labour market programs across the world to support decisions on how to allocate resources such as job training opportunities (Desiere, Langenbucher and Struyven, 2019). These systems aim to predict an individual's risk of long-term unemployment such that costly services can be targeted more efficiently, ideally to those for whom the intervention will have the greatest positive impact. Different models use different predictors. For example, gender has been explicitly included in profiling models used in Australia (Lipp, 2005) and Austria (Holl, Kernbeiß and Wagner-Pinter, 2018).

A recent example is the Austrian AMS algorithm, which uses an undisclosed stratification procedure (Gamper, Kernbeiß and Wagner-Pinter, 2020) that is approximated by a published logistic regression model (Holl, Kernbeiß and Wagner-Pinter, 2018), and includes gender as a feature. The AMS algorithm has sparked significant legal and political controversy, leading to multiple revisions and delays in its deployment. As a result, it has not yet been implemented in practice. Critics argue that the model creates a false impression of objectivity and has been specifically accused of reinforcing gender discrimination (Allhutter et al., 2020). The Austrian system stands out by targeting individuals of medium risk: the model categorizes job seekers into three groups—low (A), medium (B) and high (C) risk of long-term unemployment—but only those in the medium-risk group (B) are offered access to the costly job training programs. There are various issues with this approach, most notably its reliance on observational data to make causal claims (Sharma and Wilder, 2025; Zezulka and Genin, 2024), such as the assumption that individuals at medium risk benefit the most from the intervention. In line with the broader theme of this paper, we highlight another common but flawed assumption: that using more attributes necessarily leads to more objective (and hence fairer) decisions. One major criticism of the AMS algorithm has been its discriminatory impact, especially

11 While our analyses have focused on the 50% threshold, we show below that the underlying insights apply more broadly.

with regard to gender. The model assigned a negative coefficient of −0.14 to women and applied an additional penalty for care obligations, a variable that is only present for women. As a result, 5% of women were placed in the high-risk group C—deemed ineligible for job training—when compared to only 3% of men (Holl, Kernbeiß and Wagner-Pinter, 2018). Defenders of the system argued that its predictions simply reflect actual labour market conditions, stating it follows *"the real chances on the labour market as far as possible"* (as cited in Allhutter et al. (2020)).

TABLE 6.2: Group C sizes for both the aware and unaware LR and GBM models on the original ALMP dataset (means and STDs over 10 random splits). The Δ column reflects the difference between men and women. Note how the unaware models yield more equitable results with negligible loss in accuracy. Best results indicated in bold font.

Model	Mode	Male	Female	$\Delta \downarrow$	AUROC ↑
LR	Aware	1.6% ± .1%	2.2% ± .2%	0.7% ± .3%	**0.680** ±.002
	Unaware	1.7% ± .1%	2.0% ± .2%	**0.2%** ±.2%	**0.680** ±.002
GBM	Aware	3.9% ± .3%	8.3% ± .4%	4.4% ± .4%	**0.700** ±.003
	Unaware	4.5% ± .1%	6.1% ± .2%	**1.6**% ±.2%	0.697 ± .003

However, as shown in previous sections, there is no direct or reliable link between observed label rates (or supposed 'real chances') and the disparities in outcomes produced by model-based decisions. In the following analysis, we reconstruct the AMS setting as faithfully as possible (given the lack of access to the original data) and explore the effects of excluding gender on both accuracy and disparate impact. Lacking the original data, we base our analysis on the active labour market policy (ALMP) dataset from Switzerland (Lechner et al., 2020), focusing on the German-speaking cantons. Although the features in this dataset differ somewhat from those used in the AMS model, we aim to approximate the setting, not replicate it exactly. We adopt the same prediction target used by the AMS algorithm to determine assignment to group C, namely, whether an individual is employed for at least 6 out of the next 24 months.[12] We train both LR and GBM models and present the results in Table 6.2. In our analysis, we focus on two main outcomes: (1) the per-gender proportion of individuals assigned to group C (those considered too high-risk to receive job training), and the difference between genders; and (2) the AUROC of the models, given that the decision criterion is the 25% threshold. As reflected in Table 6.2, while omitting gender has little impact on AUROC, it leads to a more balanced assign-

12 A second model with a slightly different target is used to determine assignment to group A.

ment between genders, *i.e.*, the proportion of individuals deemed too risky is more equitable in the unaware model. Hence, even if there is no straightforward connection between AUROC and efficiency, these results suggest that Fairness through Unawareness can lead to more equitable but equally efficient policies in practice.

6.7 CONCLUSION

In this paper, we have shown that unaware models, which omit a protected attribute, can significantly reduce disparate impact across groups defined by that attribute while maintaining accuracy. Our theoretically and empirical analysis suggests that this effect is common, at least on tabular data. Our findings urge caution when using logistic regression, as it can produce unintended disparities. Note that our analyses do not suggest that LR or protected attributes should never be used (although there can be legal constraints), nor that FtU is sufficient for ensuring fair algorithmic decision-making. Indeed, we did not compare FtU to fairness interventions, such as threshold adjustments, precisely because our point is less about achieving SOTA results than about demonstrating the importance of more basic modelling choices (even though such a comparison may be of interest for the intervention perspective). In this sense, our results underscore the need for critical evaluation of algorithmic choices as part of the broader deployment pipeline of machine learning models (Black et al., 2024a). More generally, our results challenge the notion of a strict fairness-accuracy trade-off, which has been theoretically suggested based on true distributions and Bayes-optimal predictors (Corbett-Davies et al., 2017; Menon and Williamson, 2018). We argue that the relevance of such theoretical claims to real-world applications should be scrutinised, given the potential dangers of uncritically applying idealised assumptions to real-world problems with consequential impact on people's lives (Höltgen and Williamson, 2026b).

6.8 APPENDIX

6.8.1 *More details on experimental setups*

6.8.1.1 *Datasets*

The Swiss ALMP dataset features include education, demographics (gender, marital status, age), nationality, region, three related to employment history, two macroeconomic indicators, eight features on previous job characteristics,

TABLE 6.3: Features used in ACS Employment and Income.

ACS Employment	**ACS Income**
SCHL: Education	SCHL: Education
MAR: Marital status	MAR: Marital status
AGEP: Age	AGEP: Age
RAC1P: Race	RAC1P: Race
RELP: Relationship	RELP: Relationship
SEX: Sex	SEX: Sex
DEYE: Vision difficulty	
ANC: Ancestry	
DREM: Cognitive difficulty	
MIG: Move in last year	
CIT: Citizenship	
DIS: Disability	
DEAR: Hearing difficulty	
MIL: Military service	
NATIVITY: Born in US	
ESP: Employment status of parents	
	OCCP: Occupation
	WKHP: Work hours per week
	COW: Class of worker
	POBP: Place of birth

and ten features on completed job training programs. More information can be found on the dataset website (Lechner et al., 2020). The features used in the ACS datasets are listed in Table 6.3. For more details, we refer to the original publication (Ding et al., 2021).

6.8.1.2 *Models*

We use the default settings of `LogisticRegression` in `scikit-learn` but tune `GradientBoostingClassifier` for higher accuracy. The adapted parameters are `n_estimators=500`, `max_depth=10`, `max_leaf_nodes=50`, `min_samples_leaf=500`, very similar to the optimal parameters on the ACS datasets according to previous work (Cruz and Hardt, 2024, personal communication). We use the same

TABLE 6.4: Datasets summary. The number of features differs by one (protected attribute) between aware and unaware models. The 'Group Size' column indicates the proportion of data points belonging to the advantaged group, and the 'DBR' column contains the Disparate Benefit Ratio, which is the ratio of favourable outcomes received by the disadvantaged group compared to the advantaged group, indicating the degree of inequality in model predictions across protected groups. When this ratio is much smaller than 1, it indicates disparate treatment.

Dataset	# Feat	PA	PA Values	Group Size	DBR	Size
ACS Inc	9/10	Sex	male/female	52.1%	0.15	1,664,500
		Race	white/black	89.8%	0.14	1,445,699
ACS Emp	15/16	Sex	male/female	49.0%	0.06	3,236,107
		Race	white/black	88.9%	0.06	2,796,670
ALMP	28/29	Gender	male/female	56.2%	0.04	69,372
		Citizenship	Swiss/foreign	64.3%	0.17	69,372

settings for the ALMP experiments. The accuracies for the default aware models are reported in Table 6.5 for comparability.

TABLE 6.5: Accuracies per setting for the aware model with standard deviations. The accuracies of aware models on ACS datasets differ between the sex and the race setting since we only use a subset of the data in the latter, restricting to 'white' and 'black'.

Model	ACS Inc (s)	ACS Inc (r)	ACS Emp (s)	ACS Emp (r)	ALMP
LR	77.4% ± .1%	77.1% ± .0%	77.9% ± .0%	78.1% ± .1%	65.0% ± .4%
GBM	82.1% ± .1%	81.8% ± .1%	83.2% ± .1%	83.2% ± .0%	66.2% ± .3%

6.8.1.3 *Single feature experiment*

Here, we give more details on the single feature experiments described in Section 6.3.2. Classification rates for the aware model are at 0.477 (men) and 0.212 (women). Histograms of the X value distribution (Figure 6.6) show that among women, there is a higher fraction of highly educated individuals, especially with a Master's degree. This results in a higher classification rate for women (0.449) compared to men (0.389) in the unaware model, still significantly reducing the (absolute) disparate impact. We also show model predictions for aware and unaware GBM models (Figure 6.5). Here, the threshold for the unaware model is the same as for LR, whereas the aware model has a higher threshold for men

(21) compared with the aware LR model (same as the threshold of the unaware models). The accuracy of the aware GBM model is slightly higher, at 70.3%, and the classification rates are 0.389 (men) and 0.212 (women). This still means a much higher disparate impact than for the unaware model.

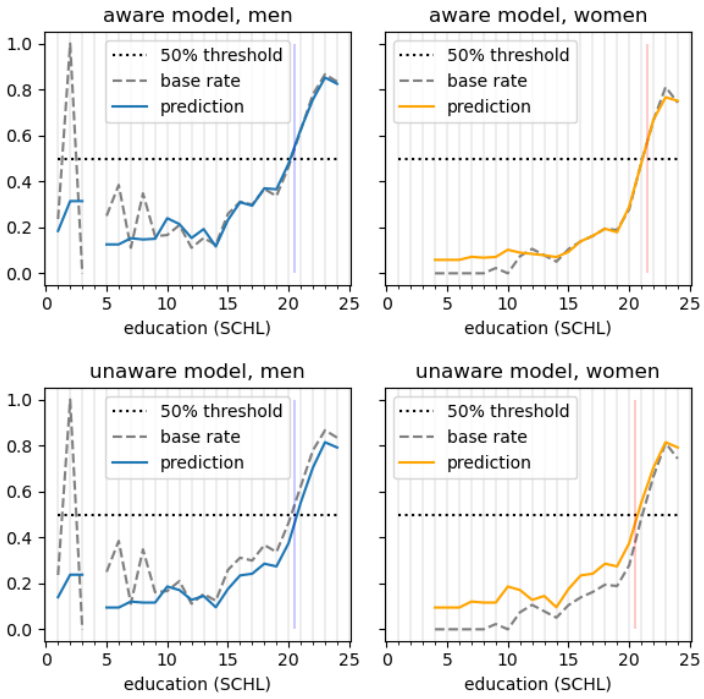

FIGURE 6.5: Predictions for aware and unaware GBM models on ACS Income NY when only using education as an input. Vertical colors indicate the 50% cutoff, highlighting that the aware GBM model (top) requires much higher education for women than for men.

6.8.2 *Proofs*

6.8.2.1 *Results in Section 6.3*

Proposition 6.2 (Relation between Base Rates and Disparate Impact)**.** *Assume* f *is calibrated as in (6.4), satisfying* $\forall v \in [0, 0.5) : P_{\mathcal{B}}(v) \geqslant P_{\mathcal{A}}(v)$ *and* $\forall v \in [0.5, 1] : P_{\mathcal{A}}(v) \geqslant P_{\mathcal{B}}(v)$*. Then for* F *defined as in (6.1),* $\mathsf{DI}(F) \geqslant \mathsf{DBR}$.

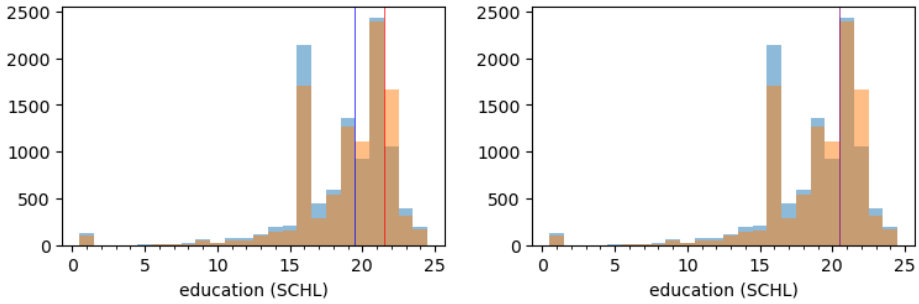

FIGURE 6.6: Histogram of the distribution of X values along the education dimension for men (blue) and women (orange/red). vertical lines denote the thresholds induced by the aware (left) and unaware (right) model.

Proof.

$$\begin{aligned}
\mathrm{DI}(F) &= P_{\mathcal{A}}(V_a \geqslant 0.5) - P_{\mathcal{B}}(V_b \geqslant 0.5) && (6.29)\\
&= \int_{[0.5,1]} P_{\mathcal{A}}(v) - P_{\mathcal{B}}(v)\ dv && (6.30)\\
&\geqslant \int_{F} (P_{\mathcal{A}}(v) - P_{\mathcal{B}}(v)) \cdot v\ dv && (6.31)\\
&\geqslant \int_{[0.5,1]} (P_{\mathcal{A}}(v) - P_{\mathcal{B}}(v)) \cdot v\ dv + \int_{[0,0.5)} (P_{\mathcal{A}}(v) - P_{\mathcal{B}}(v)) \cdot v\ dv && (6.32)\\
&= \int_{[0,0.5)} P_{\mathcal{A}}(v) \cdot v\ dv + \int_{[0.5,1]} P_{\mathcal{A}}(v) \cdot v\ dv \\
&\quad - \left(\int_{[0,0.5)} P_{\mathcal{B}}(v) \cdot v\ dv + \int_{[0.5,1]} P_{\mathcal{B}}(v) \cdot v\ dv \right) && (6.33)\\
&= \mathbb{E}_{P_{\mathcal{A}}}[V_a] - \mathbb{E}_{P_{\mathcal{B}}}[V_b] = \mathrm{DBR} && (6.34)
\end{aligned}$$

□

As can be seen from the proof, the assumptions in Proposition 6.2 can be relaxed to, respectively,

$$\int_{[0.5,1]} P_{\mathcal{A}}(v) \cdot (1-v)\ dv \geqslant \int_{[0.5,1]} P_{\mathcal{B}}(v) \cdot (1-v)\ dv, \tag{6.35}$$

and

$$\int_{[0,0.5)} P_{\mathcal{B}}(v) \cdot v\ dv \geqslant \int_{[0,0.5)} P_{\mathcal{A}}(v) \cdot v\ dv. \tag{6.36}$$

The left hand side of (6.36) is equivalent to $\mathbb{E}_{P_{\mathcal{B}}}[V_b \mid V_b < 0.5] \cdot P(\{V_b < 0.5\})$, i.e. the mean of the predictions below the decision threshold times the probability mass below the threshold.

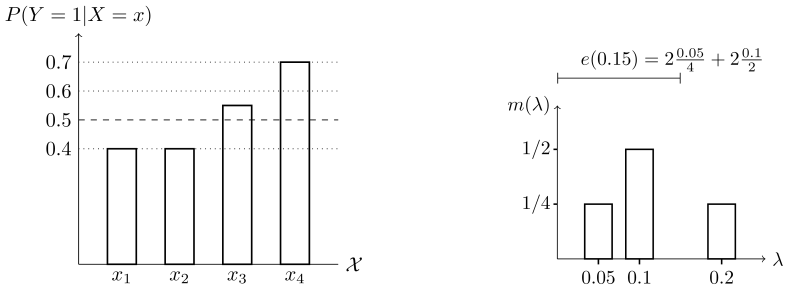

FIGURE 6.7: Illustration of $e(\lambda)$ in the case of the earlier example: $e(\lambda)$ is the additional error if for all predictions closer to 50% than λ, the suboptimal classification is taken. An illustration for $\lambda = 0.15$ is given.

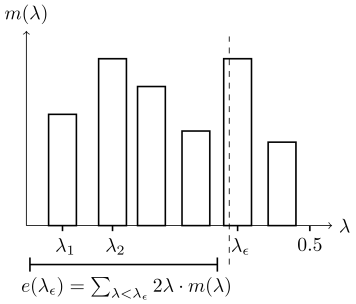

FIGURE 6.8: Illustration of λ_ϵ and $e(\lambda_\epsilon)$: λ_ϵ is the highest divergence from 0.5 s.t. switching all predictions closer than λ_ϵ to 50% increases the error rate by less than ϵ.

6.8.2.2 *Results in Section 6.4*

We start with the lemma on randomised classifiers.

Lemma 6.4.
The bound (6.46) can be achieved in any setting for every ϵ by a randomised *classifier defined as*

$$F_\epsilon(x) = \begin{cases} 1 - B(x) & \text{for } |p(x) - 0.5| < \lambda_\epsilon \\ 1 - B(x) \text{ with probability } \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon \cdot m(\lambda_\epsilon)} & \text{for } |p(x) - 0.5| = \lambda_\epsilon \\ B(x) & \text{for } |p(x) - 0.5| > \lambda_\epsilon. \end{cases} \tag{6.37}$$

Proof.
Take some $\mathcal{X}$ and P and ϵ. Then

$$\begin{aligned}
\mathrm{Acc}(F_\epsilon) = &\sum_{\lambda<\lambda_\epsilon} m(\lambda)\cdot(0.5-\lambda) + \sum_{\lambda>\lambda_\epsilon} m(\lambda)\cdot(0.5+\lambda) \\
&+ \Bigg(m(\lambda_\epsilon)\cdot\frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\cdot(0.5-\lambda_\epsilon) \\
&\qquad + m(\lambda_\epsilon)\cdot\left(1-\frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\right)\cdot(0.5+\lambda_\epsilon)\Bigg) && (6.38)\\
= &\sum_{\lambda<\lambda_\epsilon} m(\lambda)\cdot(0.5+\lambda-2\lambda) + \sum_{\lambda>\lambda_\epsilon} m(\lambda)\cdot(0.5+\lambda) \\
&+ m(\lambda_\epsilon)\cdot\frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\cdot(0.5+\lambda_\epsilon-2\lambda_\epsilon) \\
&+ m(\lambda_\epsilon)\cdot\left(1-\frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\right)\cdot(0.5+\lambda_\epsilon) && (6.39)\\
= &\sum_{\lambda<\lambda_\epsilon} m(\lambda)\cdot(0.5+\lambda) + \sum_{\lambda>\lambda_\epsilon} m(\lambda)\cdot(0.5+\lambda) - \sum_{\lambda<\lambda_\epsilon} m(\lambda)\cdot 2\lambda \\
&+ m(\lambda_\epsilon)\cdot(0.5+\lambda_\epsilon) - m(\lambda_\epsilon)\cdot\frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\cdot 2\lambda_\epsilon && (6.40)\\
= &\sum_{\lambda\in\Lambda} m(\lambda)\cdot(0.5+\lambda) - e(\lambda_\epsilon) - m(\lambda_\epsilon)\cdot\frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\cdot 2\lambda_\epsilon && (6.41)\\
= &\mathrm{Acc}(B)-\epsilon && (6.42)
\end{aligned}$$

and

$$\begin{aligned}
d(F_\epsilon, B) &= \mu(\{x\in\mathcal{X} : F(x)\neq B(x)\}) && (6.43)\\
&= \sum_{\lambda<\lambda_\epsilon} m(\lambda) + \frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon\cdot m(\lambda_\epsilon)}\cdot m(\lambda_\epsilon) && (6.44)\\
&= \frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon} + \sum_{\lambda<\lambda_\epsilon} m(\lambda). && (6.45)
\end{aligned}$$

□

This lemma helps to prove our new bound on model multiplicity.

Proposition 6.3 (Upper Bound on Multiplicity)**.**

$$\max_{F\in R_\epsilon(B)} d(F,B) \leqslant \frac{\epsilon-e(\lambda_\epsilon)}{2\lambda_\epsilon} + \sum_{\lambda<\lambda_\epsilon} m(\lambda) \tag{6.46}$$

This bound is the lowest possible bound with only access to the cumulative distribution of $p(x)$*.*

Proof.
We now show that no randomised classifier represented as $F : \mathcal{X}\to[0,1]$ with

$Acc(F)\lambda_\epsilon Acc(B) - \epsilon$ can disagree more with B on expectation than F_ϵ. Then *a forteriori*, there can be no deterministic classifier that breaks the bound. Now for any randomised classifier F with $d(F,B) > d(F_\epsilon, B)$, we show $Acc(F) < Acc(F_\epsilon)$. Let the average prediction on $U(\lambda)$ be given by

$$f(\lambda) := \frac{1}{m(\lambda)} \sum_{x \in U(\lambda)} \mu(\{x\}) |F(x) - 0.5|. \tag{6.47}$$

By construction of F_ϵ,

$$\forall \lambda < \lambda_\epsilon : \sum_{x \in U(\lambda)} |B(x) - F_\epsilon(x)| \geqslant |B(x) - F(x)| \tag{6.48}$$

and

$$\forall \lambda > \lambda_\epsilon : \sum_{x \in U(\lambda)} |B(x) - F_\epsilon(x)| \leqslant |B(x) - F(x)| \tag{6.49}$$

Then $d(F,B) > d(F_\epsilon, B)$ implies that

$$\begin{aligned} 0 &\leqslant \sum_{\lambda < \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F_\epsilon(x)| - |B(x) - F(x)|) \cdot \mu(\{x\}) && (6.50) \\ &< \sum_{\lambda \geqslant \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F(x)| - |B(x) - F_\epsilon(x)|) \cdot \mu(\{x\}) && (6.51) \end{aligned}$$

But since $\forall \lambda \in \Lambda$, $x \in U(\lambda)$ and any randomised classifier G,

$$\mathbb{E}_P[L(G(X), Y) \mid X = x] = \mathbb{E}_P[L(B(X), Y) \mid X = x] + |B(x) - G(x)| \cdot 2\lambda \tag{6.52}$$

we get

$$\begin{aligned} &Acc(F_\epsilon) - Acc(F) \\ &\quad = Acc(B) - \sum_{\lambda \in \Lambda} \sum_{x \in U(\lambda)} |B(x) - F_\epsilon(x)| \cdot 2\lambda \cdot \mu(\{x\}) \\ &\qquad - Acc(B) - \sum_{\lambda \in \Lambda} \sum_{x \in U(\lambda)} |B(x) - F(x)| \cdot 2\lambda \cdot \mu(\{x\}) && (6.53) \\ &\quad = \sum_{\lambda \in \Lambda} \sum_{x \in U(\lambda)} (|B(x) - F(x)| - |B(x) - F_\epsilon(x)|) \cdot 2\lambda \cdot \mu(\{x\}) && (6.54) \\ &\quad = \sum_{\lambda \geqslant \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F(x)| - |B(x) - F_\epsilon(x)|) \cdot 2\lambda \cdot \mu(\{x\}) \\ &\qquad - \sum_{\lambda < \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F_\epsilon(x)| - |B(x) - F(x)|) \cdot 2\lambda \cdot \mu(\{x\}) && (6.55) \\ &\quad \geqslant \sum_{\lambda \geqslant \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F(x)| - |B(x) - F_\epsilon(x)|) \cdot 2\lambda_\epsilon \cdot \mu(\{x\}) \\ &\qquad - \sum_{\lambda < \lambda_\epsilon} \sum_{x \in U(\lambda)} (|B(x) - F_\epsilon(x)| - |B(x) - F(x)|) \cdot 2\lambda_\epsilon \cdot \mu(\{x\}) && (6.56) \\ &\quad > 0. && (6.57) \end{aligned}$$

We now show that the bound is tight in the sense that for any cumulative distribution (defined by Λ and m), we can i) for each ϵ find a $\mathcal{X}$, $P(X)$ and ii) for each $\mathcal{X}$, $P(X)$ find an $\epsilon > 0$ so that the bound is achievable.

Take some Λ, m and ϵ. Then let $\mathcal{X}$ be such that there is $x_0 \in U(\lambda_\epsilon)$ with $\mu(\{x_0\}) = \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon}$ and let

$$F(x) = \begin{cases} 1 - B(x) & \text{for } |p(x) - 0.5| < \lambda_\epsilon \\ 1 - B(x) & \text{for } x = x_0 \\ B(x) & \text{otherwise.} \end{cases} \tag{6.58}$$

Then, similar to F_ϵ in the Lemma,

$$\begin{aligned} \mathrm{Acc}(B) - \mathrm{Acc}(F_\epsilon) &= \sum_{\lambda \in \Lambda} \sum_{x \in U(\lambda)} |B(x) - F_\epsilon(x)| \cdot 2\lambda \cdot \mu(\{x\}) && (6.59)\\ &= 2\lambda_\epsilon \cdot \mu(\{x_0\}) + \sum_{\lambda < \lambda_\epsilon} 2\lambda \cdot m(\lambda) && (6.60)\\ &= 2\lambda_\epsilon \cdot \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon} + e(\lambda_\epsilon) && (6.61)\\ &= \epsilon && (6.62) \end{aligned}$$

and

$$\begin{aligned} d(F, B) &= \mu(\{x \in \mathcal{X} : F(x) \neq B(x)\}) && (6.63)\\ &= \mu(\{x\}) + \sum_{\lambda < \lambda_\epsilon} m(\lambda) && (6.64)\\ &= \frac{\epsilon - e(\lambda_\epsilon)}{2\lambda_\epsilon} + \sum_{\lambda < \lambda_\epsilon} m(\lambda). && (6.65) \end{aligned}$$

□

	x_1	x_2	x_3	x_4
$B(x)$	0	0	1	1
$F_1(x)$	0	0	0	1
$F_2(x)$	0	1	0	1
$F_3(x)$	1	1	1	1
$F_4(x)$	1	1	0	1

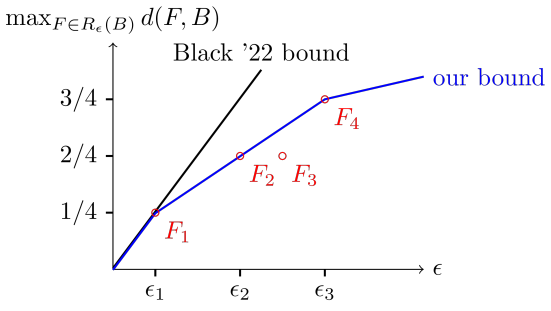

FIGURE 6.9: Illustration of Proposition 6.3. The table gives the classifiers derived from the calibrated predictors in Table 6.1 plus an additional classifier F_4. We show that our bound refines the bound (6.19) from Black, Raghavan and Barocas (2022), and is applicable under on fewer assumptions.

Corollary 6.5.
Under assumption (6.20),

$$\max_{F \in R_\epsilon(B)} d(F, B) \leqslant \frac{\epsilon}{2c}. \tag{6.66}$$

Proof.
From (6.20), we get $\forall \lambda \in \Lambda : \lambda \geqslant c$ and hence

$$\frac{\epsilon - e(\lambda_\epsilon)}{2\lambda} + \sum_{\lambda < \lambda_\epsilon} m(\lambda) \leqslant \frac{\epsilon - e(\lambda_\epsilon)}{2c} + \sum_{\lambda < \lambda_\epsilon} m(\lambda) \frac{2\lambda}{2c} = \frac{\epsilon - e(\lambda_\epsilon)}{2c} + \frac{e(\lambda_\epsilon)}{2c} = \frac{\epsilon}{2c} \tag{6.67}$$

and thus the Proposition entails

$$\max_{F \in R_\epsilon(B)} d(F, B) \leqslant \frac{\epsilon}{2c}. \tag{6.68}$$

□

Proposition 6.6 (Multiplicity Bound for Non-optimal Models)**.**
We can bound the disagreement between two classifiers $F \in R_\epsilon(B)$ *and* $G \in R_\delta(B)$ *via*

$$d(F, G) \leqslant \max_{H \in R_{\epsilon+\delta}(B)} d(H, B). \tag{6.69}$$

Proof.
Fix F and G and let $U := \{x \in \mathcal{X} : F(x) \neq G(x)\}$. Now define a classifier H via

$$H(x) := \begin{cases} 1 - B(x) & \text{for } x \in U \\ B(x) & \text{for } x \notin U. \end{cases} \tag{6.70}$$

Then $d(B, H) = \mu(U) = d(F, G) \leqslant d(F, B) + d(G, B) \leqslant \epsilon + \delta$ and thus $H \in R_{\epsilon+\delta}(B)$. The result follows. □

Proposition 6.7 (Upper Bound on Disparate Impact)**.**
For two classifiers F *and* G*, the difference in their disparate impact can be bounded based on their disagreement and the size of the smaller group:*

$$|DI(F) - DI(G)| \leqslant \frac{d(F, G)}{\min_{* \in \{\mathcal{A}, \mathcal{B}\}} P(X \in *)}. \tag{6.71}$$

Proof.

$$\begin{aligned} |DI(F) - DI(G)| &= |P(F(X) = 1 \mid X \in \mathcal{A}) - P(F(X) = 1 \mid X \in \mathcal{B}) \\ &\quad - P(G(X) = 1 \mid X \in \mathcal{A}) + P(G(X) = 1 \mid X \in \mathcal{B})| && (6.72) \\ &= |P(F(X) = 1 \mid X \in \mathcal{A}) - P(G(X) = 1 \mid X \in \mathcal{A}) \\ &\quad + P(G(X) = 1 \mid X \in \mathcal{B}) - P(F(X) = 1 \mid X \in \mathcal{B})| && (6.73) \\ &\leqslant P(F(X) \neq G(X) \mid X \in \mathcal{A}) + P(F(X) \neq G(X) \mid X \in \mathcal{B}) && (6.74) \\ &\leqslant d(F, G) \,/\, \min_{* \in \{\mathcal{A}, \mathcal{B}\}} P(X \in *) && (6.75) \end{aligned}$$

since

$$\begin{aligned} d(F,G) &= P(F(X) \neq G(X)) && (6.76)\\ &= P(F(X) \neq G(X) \mid X \in \mathcal{A}) \cdot P(X \in \mathcal{A}) \\ &\quad + P(F(X) \neq G(X) \mid X \in \mathcal{B}) \cdot P(X \in \mathcal{B}) && (6.77)\\ &\geqslant P(F(X) \neq G(X) \mid X \in \mathcal{A}) \cdot \min_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *) \\ &\quad + P(F(X) \neq G(X) \mid X \in \mathcal{B}) \cdot \min_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *). && (6.78) \end{aligned}$$

□

Proposition 6.8 (Achievable DI for Unaware Models)**.**
Fix P*, an unaware model* F*, and some* $\epsilon > 0$*. Then there is a model* $G \in R_\epsilon(F)$ *with*

$$|DI(F) - DI(G)| \geqslant \frac{1}{2 \cdot \max_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *)} \max_{H \in R_{\epsilon+\delta}(B)} d(H,B). \tag{6.79}$$

Proof.
Fix an H that achieves the maximum and let $U := \{x \in \mathcal{X} : H(x) \neq B(x)\}$. Then define two classifiers $G_>$ and $G_<$ via

$$G_>(x) := \begin{cases} 1 & \text{for } x \in U \cap \mathcal{A} \text{ and } F(x) = 0 \\ 0 & \text{for } x \in U \cap \mathcal{B} \text{ and } F(x) = 1 \\ F(x) & \text{otherwise,} \end{cases} \tag{6.80}$$

and

$$G_<(x) := \begin{cases} 1 & \text{for } x \in U \cap \mathcal{B} \text{ and } F(x) = 0 \\ 0 & \text{for } x \in U \cap \mathcal{A} \text{ and } F(x) = 1 \\ F(x) & \text{otherwise.} \end{cases} \tag{6.81}$$

Then

$$\{x \in \mathcal{X} : H(x) \neq B(x)\} = U = \{x \in \mathcal{X} : F(x) \neq G_<(x)\} \cup \{x \in \mathcal{X} : F(x) \neq G_>(x)\} \tag{6.82}$$

and hence

$$d(F, G_\#) \geqslant \frac{1}{2} d(H,B) \tag{6.83}$$

for one $\# \in \{<, >\}$. Furthermore, we get

$$|DI(F) - DI(G_\#)| \geqslant \frac{d(F, G_\#)}{\max_{*\in\{\mathcal{A},\mathcal{B}\}} P(X \in *)} \tag{6.84}$$

by the same derivation as for (6.71) but with equality in (6.74) due to the construction of $G_\#$ and with '$\geqslant$' in (6.75) by switching min to max. □

Proposition 6.9 (Lower Bound on DI for LR)**.**
Let $F : \mathcal{X} \to \{0, 1\}$ *be an PA-aware classifier based on a LR model* f *with coefficient* $c_G < 0$ *for the PA, i.e. the disadvantaged group gets a pre-sigmoid penalty of* c_G*. Further, let*

$$Q(c_G) := \{x \in \mathcal{B} : f(x) \in [\sigma(c_G), 0.5]\} \tag{6.85}$$

denote the set of points that get predictions between $\sigma(c_G)$ *and* 0.5 *by* f and *are in Group* b*. Then the corresponding unaware classifier* F' *based on setting* $c_{PA} \leftarrow 0$ *in* F *has accuracy bounded by*

$$|\mathrm{Acc}(F) - \mathrm{Acc}(F')| \leqslant 2\sigma(-c_G) \cdot P(X \in Q(c_G)) \tag{6.86}$$

and disparate impact given by

$$\mathrm{DI}(F') = \mathrm{DI}(F) - P(X \in Q(c_G))/P(X \in \mathcal{B}) \tag{6.87}$$

if group b *is still disadvantaged in* F' *(i.e.* $\mathrm{DI}(F) \geqslant P(X \in Q(c_G))/P(X \in \mathcal{B})$*), and otherwise*

$$|\mathrm{DI}(F) - \mathrm{DI}(F')| \leqslant P(X \in Q(c_G))/P(X \in \mathcal{B}). \tag{6.88}$$

Proof.
F' and F only differ on the datapoints that are classified as 0 by F but as 1 by F'. These are the datapoints that have prediction < 0.5 by f but $\geqslant 0.5$ by f', i.e.

$$\{x \in \mathcal{X} : \sigma^{-1}(f(x)) < 0 \wedge \sigma^{-1}(f'(x)) \geqslant 0\} = \{x \in \mathcal{B} : -c_G \leqslant \sigma^{-1}(f(x)) < 0\}. \tag{6.89}$$

Now this set of points is $Q(c_G)$, and the maximum change in accuracy occurs if one model is right on all points, whereas the other model is wrong on all points, *and* all predictions are as extreme as possible. Then, assuming that one of the models is calibrated on that set, the label ratio is bounded from both sides by $\sigma(-c_G) \leqslant P(Y \mid X \in Q(c_G)) \leqslant \sigma(c_G)$. This means the change in accuracy is bounded by $2 \cdot \sigma(c_G) \cdot P(X \in Q(c_G))$, giving us (6.86.

Recall that disparate impact is defined as

$$\mathrm{DI}(F) = |P(F(X) = 1 \mid X \in \mathcal{A}) - P(F(X) = 1 \mid X \in \mathcal{B})|. \tag{6.90}$$

Now only predictions on group b are changed, such that

$$P(F(X) = 1 \mid X \in \mathcal{A}) = P(F'(X) = 1 \mid X \in \mathcal{A}). \tag{6.91}$$

On the other hand, we have

$$P(F'(X) = 1 \mid X \in \mathcal{B}) = P(\{F(X) = 1 \vee -c_G \leqslant \sigma^{-1}(f(x)) < 0 \mid X \in \mathcal{B}\}), \tag{6.92}$$

and thus

$$P(F'(X) = 1 \mid X \in \mathcal{B}) = P(F(X) = 1 \mid X \in \mathcal{B}) + P(X \in Q(c_G))/P(X \in \mathcal{B}). \quad (6.93)$$

Putting things together, we get

$$\begin{aligned} &\left|P(F'(X) = 1 \mid X \in \mathcal{A}) - P(F'(X) = 1 \mid X \in \mathcal{B})\right| && (6.94)\\ &\quad = |P(F(X) = 1 \mid X \in \mathcal{A}) - P(F(X) = 1 \mid X \in \mathcal{B})| - P(X \in Q(c_G))/P(X \in \mathcal{B}) && (6.95)\end{aligned}$$

if the terms inside the vertical bars remain positive. Otherwise, we still get (6.88). □

ACKNOWLEDGEMENTS

We would like to thank Adrián Arnaiz-Rodríguez for insightful discussions throughout the project and Sebastian Zezulka as well as the anonymous reviewers for helpful feedback on the manuscript.

BH is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. BH also acknowledges support by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and travel support from ELIAS (GA no 101120237).

NO is supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación) and by EU - HE ELIAS – Grant Agreement 101120237. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA).

7

POSITION: THE CATEGORIZATION OF RACE IN ML IS A FLAWED PREMISE

This is business; no faces, just lines and statistics
From your phone, your Zip Code, to SSI digits
The system break man, child, and women into figures
Two columns for 'Who is' and 'Who ain't' [black]
Numbers is hard and real and they never have feelings
But you push too hard, even numbers got limits

— Yasiin Bey, *Mathematics*

AUTHOR CONTRIBUTIONS

The original idea of the project is due to Miriam Doh and Piera Riccio and was later refined in discussions with Benedikt Höltgen. The empirical analysis was done by Miriam Doh. Most of the conceptualisation and writing of the final paper are due to Miriam Doh and Benedikt Höltgen, in close collaboration with Piera Riccio, who helped with both structuring and writing the paper. The project was supervised by Nuria Oliver, including valuable feedback on the writing.

ABSTRACT

This position paper critiques the reliance on rigid racial taxonomies in machine learning, exposing their U.S.-centric nature and lack of global applicability—particularly in Europe, where race categories are not commonly used. These classifications oversimplify racial identity, erasing the experiences of mixed-race individuals and reinforcing outdated essentialist views that contradict the social construction of race. We suggest research agendas in machine learning that move beyond categorical variables to better address discrimination and social inequality.

7.1 INTRODUCTION

> "A race is only a sort of average of a large number of individuals; and averages differ from one another much less than individuals. Popular impression exaggerates the differences, accurate measurements reduce them" (Kroeber, 1948)

The concept of separate human races arose in the 17th and 18th centuries and was used by Westerners to justify slavery despite their Christian faith (Smedley and Smedley, 2005). This notion persisted into the 20th century and was closely tied to early statistics and eugenics, with pioneers like Galton, Pearson, and Fisher reinforcing the idea of biological racial separability. However, anthropologists have increasingly contested these racial distinctions, recognizing them as social constructs rather than biologically discrete entities (Kroeber, 1948). The "one-drop rule" in the U.S. and the Apartheid regime's classification system in South Africa illustrate how racial categories have been historically fluid and politically motivated (Bowker and Star, 1999).

For decades, scholars across disciplines have emphasized that racial classifications are neither genetically discrete (Wilson et al., 2001) nor can they be reliably measured or considered scientifically meaningful (Smedley and Smedley, 2005). Instead, race is understood as a construct whose meaning is constituted by social arrangements, practices, and intersubjective beliefs (Hu and Kohler-Hausmann, 2020). This understanding implies that there is no "correct" racial taxonomy derived from biology, as perceptions of race depend on both phenotypic traits and contextual interpretations.

Despite this knowledge, the Machine Learning (ML) research community has often relied on datasets that label race as a categorical variable, treating it as a "ground-truth" that simplifies its social and historical complexity (Abdu, Pasquetto and Jacobs, 2023). Racial labels such as "White," "Black," or "Asian" are indeed widely used in image (Karkkainen and Joo, 2021; Phillips et al., 2000; Ricanek and Tesafaye, 2006; Zhang, Song and Qi, 2017) and tabular (Angwin et al., 2016; Kohavi, 1996) datasets, often adopting U.S. census-based racial classifications without contextualization, reinforcing North American constructs in global ML research.

By using these categorizations, ML models fail to account for the social, cultural and historical contexts that shape racial identities. Even when these labels are well-intended and used for fairness audits (Ravishankar et al., 2023; Wang and Deng, 2020; Yucer et al., 2024b), they reflect outdated racial theories. This lack of critique has real-world consequences: from biased hiring algorithms to unequal healthcare diagnostics, these systems risk exacerbating disparities

rather than addressing them. To move forward, we contend that the ML community should examine how racial data is collected, labeled and used, avoiding the instrumentalization of ethical concepts, to the point that "bias becomes a form of numerical error to be corrected with better datasets, and ethics becomes a bureaucratic checklist to be inserted into the production flowchart" (Hong, 2023).

In this paper, we discuss the challenges and ethical concerns associated with the use of categorical race labels in ML systems. Thus, discussing the benefits for individuals to freely define their identities through racial categories (Rich, 2013; Stock et al., 2018) is out of its scope. Instead, we critically analyze ML practices, datasets, and the socio-historical construction of race (Section 7.2), emphasizing the dominance of U.S.-centric frameworks and their limitations from the perspective of Europe (Section 7.3) and the "mixed race" community (Section 7.4). In addition, we address the broader issue of racial reification and stereotyping in ML (Section 7.5). Finally, we propose a research agenda that avoids categorical race labels by integrating context and domain knowledge to more effectively tackle discrimination and social inequality (Section 7.6).

In summary, we argue that racial categories should be abandoned in ML research whenever possible, as they reinforce outdated essentialist views and fail to account for the complexity of human identities. We do not argue that racial discrimination should be ignored—on the contrary, we contend that fairness frameworks in machine learning must be reimagined in ways that effectively address bias without reinforcing the very inequalities they aim to mitigate.

7.2 RELATED WORK

Several studies have criticized the use of fixed racial categories in AI from different perspectives. First, the process of defining racial taxonomies has been questioned, calling for more careful and contextual choices, justification, and documentation (Abdu, Pasquetto and Jacobs, 2023; Crawford and Paglen, 2021; Gebru et al., 2021; Khan and Fu, 2021; Mickel, 2024). Second, the use of racial categories in assessing fairness and discrimination in ML has been challenged, calling for the use of different categories (Belitz et al., 2023; Benthall and Haynes, 2019; Hanna et al., 2020; Jaime and Kern, 2024). In contrast, we argue that methods aimed at measuring discrimination should go beyond using a categorical race variable.

Recent work has presented alternatives to categorical race labels both in vision and tabular data. In computer vision, the use of visual cues instead of race

has been proposed (Buolamwini and Gebru, 2018), almost exclusively focusing on skin tone (Ajmal et al., 2021; Bloomberg, 2023; Currie et al., 2024; Groh et al., 2022). We argue that other phenotypic features should be considered while avoiding "racial phenotypes" (Yucer et al., 2024a, 2022), which run the risk of reifying race along phenotypic lines (Crawford and Paglen, 2021; Hanna et al., 2020). In the case of tabular data, the most closely related work to ours concerns subgroup fairness (Kearns et al., 2018) and multi-calibration (Hébert-Johnson et al., 2018). However, the former approach still relies on predefined race categories while the latter solely aims at general predictive quality without incorporating context or distinguishing between features.

7.3 RACE TAXONOMIES IN MACHINE LEARNING: US CENTRISM VS THE EUROPEAN PERSPECTIVE

Race is often invoked as a protected attribute in the context of algorithmic fairness in ML, building on U.S. discrimination doctrines of disparate treatment or impact (Barocas and Selbst, 2016). This area has been shaped by seminal works, such as the 2016 ProPublica analysis of the COMPAS recidivism prediction algorithm, reporting higher error rates for black than for white inmates (Angwin et al., 2016). For the past decade, numerous authors have argued for fairness metrics defined in terms of comparisons between socially relevant groups, such as equal error rates or equal outcomes, and ways to optimize for them (Barocas, Hardt and Narayanan, 2023). In these works, race is typically considered a categorical attribute as any other, often adopting the taxonomy of the U.S. census, as we discuss below. Race has also been extensively used in computer vision, where many datasets with human faces contain race labels that are used to compare error rates across different racial groups (Garcia et al., 2019), provide a measure of dataset diversity (Chen, 2020), or as target variables for models serving a variety of purposes, including "security and defense, surveillance, human computer interface (HCI), biometric-based identification" (Fu, He and Hou, 2014).

Recent studies highlight how fairness research continues to rely on predefined racial classifications without critical engagement. Abdu et al. analyzed 60 ACM FAccT (2018-2020) papers (Abdu, Pasquetto and Jacobs, 2023), showing that most use racial categories without justification, primarily adopting labels from pre-existing datasets. Extending this analysis, we reviewed 78 ICML and CVPR papers from 2023 and 2024, selecting those with "fair" in the title for ICML and "fair"/"bias" for CVPR, as the latter yielded too few results when searching only for "fair".

Among these 78 papers, 45 explicitly discussed group fairness. Within this subset, 29 relied on racial categories as a protected attribute, all adopting rigid racial classifications inherited from existing datasets—confirming the persistence of these taxonomies in fairness research (Abdu, Pasquetto and Jacobs, 2023). This reliance suggests a continued adoption of established classification schemes without a critical reassessment of their validity. To further investigate this trend, we examined the most widely used ML datasets based on citation counts from Google Scholar and their recurrence in our paper analysis. Our findings confirm that these datasets overwhelmingly adopt racial classifications derived from U.S. census categories (Appendix 7.9.2), a pattern also highlighted in Abdu, Pasquetto and Jacobs (2023). Even when explicit census labels are absent, the underlying racial categories often reflect U.S.-centric taxonomies, reinforcing rigid classification structures. Notably, none of these datasets provide critical reflection of their contextual meaning or applicability beyond the United States (Benthall and Haynes, 2019; Khan and Fu, 2021; Mickel, 2024) (see Appendix 7.9.3).

Despite the cultural vicinity, the legal and institutional framework in Europe contrasts sharply with the U.S. approach, where racial categories are routinely collected and integrated into governance, research, and AI datasets. Mainly as a result of post-World War II efforts to combat racialization and discrimination, race is widely rejected as a legitimate classification category in Europe (Osanami Törngren, 2020; Rodríguez-García et al., 2021; Simon, 2017). Also at the legislative level, the European Union has institutionalized its avoidance of racial classification. For instance, the General Data Protection Regulation (GDPR) (Union, 2016) restricts the collection of racial or ethnic data, permitting it only under narrowly defined conditions, such as explicit consent or when serving a significant public interest. In practice, many European states remain hesitant to engage with race as an analytical category, relying instead on indirect indicators, such as nationality, language spoken at home, or parental country of birth (Bereni, Epstein and Torres, 2021; European Commission, 2017; Osanami Törngren, Irastorza and Rodríguez-García, 2021; Simon, 2015; Westin, 2003). Indeed, the Court of Justice of the European Union ruled that 'ethnic origin cannot be determined on the basis of a single criterion' and judicial decisions on racial or ethnic discrimination need to consider multiple factors (CJEU, 2017). Furthermore, different member states have different standards on which attributes are relevant for racial discrimination (Commission et al., 2024), highlighting the importance of local context.

As AI systems become increasingly regulated under frameworks such as the EU AI Act (Commission, 2021), with explicit requirements regarding al-

gorithmic fairness and non-discrimination, the direct adoption of U.S.-based frameworks raises significant questions (Engler, 2023; Wachter, Mittelstadt and Russell, 2021). However, even if the institutional approach to dealing with racial categories is more permissive in the United States, criticism of such taxonomies has not been exclusive to Europe. Even in the U.S., racial beliefs have long been said to "constitute myths about the diversity in the human species and about the abilities and behavior of people homogenized into "racial" categories" (American Anthropological Association, 1998).

In line with this, we argue that racial categories should be abandoned in ML research whenever possible, incorporating context and domain knowledge instead to better address discrimination. In the following sections, we support this position by identifying two problems which highlight why racial categorization in ML can be both conceptually flawed and harmful. We start by demonstrating that ML frameworks fail to represent mixed-race identities and enforce reductive classifications.

7.4 THE MIXED-RACE PROBLEM IN ML

Mixed-race individuals, defined as those with parents from different racial backgrounds or who belong to multiple racial groups (Oxford English Dictionary, 2025a), embody the complexity and fluidity of racial identities. These individuals navigate multiple cultural and racial contexts, resisting singular classification and exposing the limitations of existing taxonomies (Camara, 2016; McWatt, 2020; Remedios and Chasteen, 2013). Their dual positioning—simultaneously within and beyond established categories—makes them emblematic of identities that defy categorical frameworks (Ali, 2020). At the heart of this experience lies the concept of "in-betweenness", which highlights both the richness of multiplicity and the challenges of fragmented affiliations (Brocket, 2020; Floro, 2018). This liminal space often amplifies marginalization, reflecting how societal and institutional structures struggle to accommodate fluid identities (Ali, 2020; Camara, 2016; Nilipour, 2021).

A fundamental question emerges: *how should mixed-race identities be classified within categorical race taxonomies in AI?* There are arguably five possible ways to handle mixed-race categories in categorical race taxonomies, which we will discuss as approaches (A) to (E) below, with (A) to (C) being used in practice (Abdu, Pasquetto and Jacobs, 2023).

Approach (A) assumes that each person only belongs to one race, hence ignoring the reality of mixed-race individuals and invalidating their identity by forcing them to claim only one aspect of it (Ford, Patterson and Johnston-Guerrero,

2021; Miles, 2020). This approach is largely driven by two factors: (1) the preference for mutually exclusive racial classification schemes, which streamline computational modeling at the expense of identity fluidity, and (2) the push for computational efficiency, where minimizing the number of racial categories simplifies both the analysis and fairness auditing (Abdu, Pasquetto and Jacobs, 2023; Mickel, 2024). However, these practical considerations come at the cost of inclusivity and representational accuracy, ultimately reinforcing reductive racial frameworks rather than challenging them. This seems to be the most common approach in ML and especially in computer vision, as exemplified by the `FairFace` (Karkkainen and Joo, 2021) dataset. Marketed as a "fair" dataset, it was designed to mitigate the racial imbalances in existing face datasets by introducing seven racial groups. The authors explicitly frame their motivation as follows: *"Existing public face datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented."* This illustrates that under the assumption of rigid racial taxonomies, even well-intended approaches can become highly exclusive.

Approach (B) includes a single "Mixed-race" category, as it is the case of the American Community Survey (ACS) Public Use Microdata Sample (PUMS) data (United States Census Bureau, n.d.), from which the `folktables` dataset (Ding et al., 2021) is derived. This choice may seem to be an efficient solution given that, in the U.S. for instance, many mixed-race Black/White individuals identify as Black (and are treated as such) since the early 20th century (Thompson, 2013), and hence the mixed-race category only comprised 2.9% of U.S. census respondents as of 2010. In 2020, however, this percentage increased to 10.2%, presumably "largely due to the improvements to the design [and processing]" of the survey (United States Census Bureau, 2021).[1] However, this approach is also highly reductive: Why should, for instance, "Asian-White" be grouped together with "Black-Hispanic"?

We illustrate this limitation with an example in computer vision, using embeddings extracted with CLIP ViT-B/32 (Radford et al., 2021) from the `Chicago Face Database` (`CFD`) and its extension `CFD-MR` (Ma, Correll and Wittenbrink, 2015; Ma, Kantner and Wittenbrink, 2021). `CFD` consists of images of 597 unique individuals with their self-reported race (Asian, Black, Latino, and White), and `CFD-MR` includes images of 88 unique individuals who self-reported multi-racial ancestry (Mixed-Race). We conduct an experiment comparing the images of faces in the Asian, Black, White and Mixed-Race groups, ensuring balance across gender and sample size. As shown in Figure 7.1, we observe signific-

1 This dependence on framing as well as the fact that indicating multiple races was only possible after long negotiations in the 90s (Robbin, 2000) exemplifies the volatility of race taxonomies.

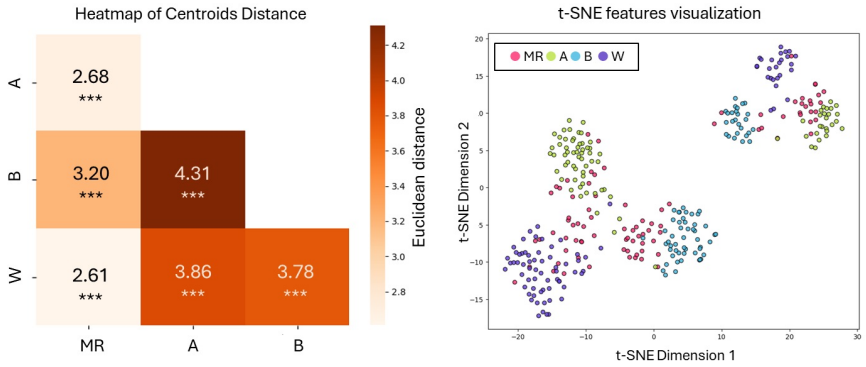

FIGURE 7.1: Embedding visualization for four racial groups from the Chicago Face Dataset and its extension, CFD-MR: Mixed-Race (MR), Asian (A), Black (B), and White (W). Left: heatmap of the Euclidean distances between group centroids, with significance: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$). Right: t-SNE plot of the embeddings of each individual image. Note how Mixed-Race individuals (depicted in magenta) occupy an intermediate space between other groups.

ant overlap between Mixed-Race samples (shown as magenta dots in the right-hand side of the Figure) and other racial groups. This blurring of the boundaries between groups in the case of mixed-race individuals further illustrates the limitation of relying on rigid racial taxonomies.

We acknowledge that CFD-MR's small sample size limits statistical robustness; however, to our knowledge, it is the only publicly available dataset including self-reported mixed-race identity, enabling analysis of mixed-race complexity without imposing external labels. The scarcity of larger, self-reported mixed-race datasets highlights a gap in available data and a shortfall in current collection practices—underscoring the urgent need for more nuanced, self-defined datasets.

Approach (C) subsumes mixed-race under "Other", as exemplified by (Xian, Yin and Zhao, 2023). This option not only inherits the limitations of using a single "Mixed-Race" category but even exacerbates them by adding to all possible variations of mixed-race individuals any other individuals that do not fit into the imposed categories.

In addition to these three approaches, we describe below two potential approaches that have not been pursued yet in the ML literature.

Approach (D) accounts for all mixed categories separately, *e.g.*, "Black/White", "Black/Hispanic", etc. However, this approach does not seem advisable for two reasons. First, it would lead to a combinatorial explosion, as for k categories, this would lead to $2^k - 1$ possible labels. Second, the categorical nature of the

variable would mean that, for instance, “Black/Hispanic" would stand in the same relationship to “Black" as to “White", which is clearly sub-optimal.

Approach (E) proposes to allow individuals to have multiple racial labels instead of being forced into a single category. While this better reflects the reality of mixed-race identities, it is not used in practice due to both data availability issues and the technical complexity of multi-label analysis. For instance, while the U.S. Census initially records race data in this way, it is later simplified into single-label categories in the ACS PUMS database, and hence also `folktables` (approach (B)). Even if such data were available, implementing this approach in ML would require new methods to handle overlapping groups, which most fairness frameworks do not support.

Rather than attempting to develop new technical solutions for such categorical approaches, we argue that race labels should be avoided altogether. We provide further justification for this position in the following section.

7.5 FROM REIFYING TO STEREOTYPING RACE IN AI

The concept of reification refers to the process of transforming abstract ideas into concrete, seemingly natural entities (Cambridge Dictionary, 2025). In the context of AI, racial reification occurs when socially constructed racial categories are encoded as computer-readable attributes and, thus, appear to be measurable and biologically real. This process not only replicates racial classification systems but entrenches race within digital infrastructures. It thereby reinforces the false notion that racial groups are stable, objective, and biologically determined rather than historically contingent and socially constructed (Hanna et al., 2020; Khan and Fu, 2021; Monea, 2019; Yucer et al., 2024b). These mechanisms, deeply embedded in machine learning systems, mirror historical physiognomic classification, where racial typologies were used to categorize, rank, and regulate human difference (Benjamin, 2019; Moss, 2021; Poole, 2005; Zhao et al., 2024).

A key issue in AI’s reification of race lies in how image datasets obtain the labels for racial categories. Many widely used datasets, such as `FairFace` (Karkkainen and Joo, 2021), rely on manual annotation, where individuals classify faces into racial categories without any contextual information. These categorizations, often based on a subjective visual assessment, do not account for the fluidity of racial identity, regional variations, or self-identification, yet they serve as the foundation for how ML models “learn" to recognize or consider race (Khan and Fu, 2021). In addition, these datasets impose a singular, racialized gaze that fixes the representation of marginalized identities into stereo-

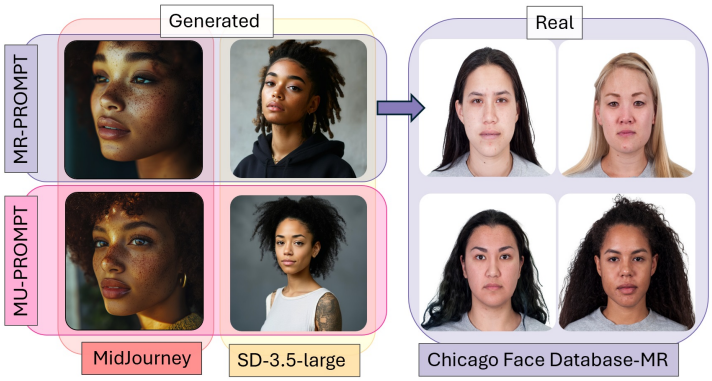

FIGURE 7.2: Comparison of AI-generated and real-world images of mixed-race individuals. The generated images come from two AI models, MidJourney and Stable Diffusion 3.5 Large, using different prompts: *"A mulatto person"* (MU-PROMPT) and *"a mixed-race person"* (MR-PROMPT). The real-world images are sourced from the Chicago Face Database-MR, showcasing greater phenotypic diversity.

typical, externally imposed frameworks. A clear example of this can be found in `ImageNet` (Deng et al., 2009), where the category "Black person" was represented in part by images of individuals performing *blackface*, illustrating how racialized bodies have historically been captured through a *white gaze* that dissects, categorizes, and fixes them in structures of power, rather than allowing for self-representation (Crawford and Paglen, 2021; Monea, 2019).

While ML-based classification systems reinforce racial categories on a large scale, generative systems for images, videos, and even audio push this further by enabling the creation of content tied to racial classifications (Bianchi et al., 2023; Ghosh, Lutz and Caliskan, 2024; Luccioni et al., 2024; Zhou et al., 2024). Unlike static datasets which have a finite number of datapoints, Generative AI models can synthesize entirely new datasets, reinforcing and expanding racial stereotypes with no human verification or historical grounding.

To illustrate this phenomenon, we conducted an experiment with Stable Diffusion 3.5 Large (Huggingface, 2024), Midjourney (MidJourney, 2024), and DALL·E 3 (OpenAI, 2024), using the prompts "a mixed-race person" (MR-PROMPT) and "a mulatto[2] person" (MU-PROMPT). The inclusion of the term mulatto in this study was intentional, as it historically frames mixed-race identity within a rigid "Black"/"White" binary (Oxford English Dictionary, 2025b). We did not include prompts such as "Asian-White," which lack an equally established cultural reference and could introduce arbitrary assumptions about how these identities should appear.

2 Disclaimer: The term *mulatto* is considered offensive in contemporary usage. It is used here solely for historical and analytical purposes to critically examine biases in Generative AI representations.

Because Generative AI systems are trained on datasets reflecting historical and cultural biases, we hypothesized that they might replicate or reinforce such reductive categorizations. Consequently, while our conclusions are necessarily limited to the specific historical stereotypes associated with these terms, this focus ensures that our findings about the replication of pre-existing biases remain clearly interpretable.

To empirically validate this hypothesis, we then generated images using each model under both prompts. Specifically, 30 images per prompt were generated for both Stable Diffusion and Midjourney. Note that we were unable to create images with DALL·E 3 due to prompt moderation on the term “mulatto". Interestingly, both models produced nearly identical images across these prompts: individuals with medium-brown skin, curly or afro-textured hair, and phenotypic traits associated with a “Black"/“White" racial mix. Figure 7.2, left, shows an example of the generated images for each prompt and model. From a quantitative perspective, the cosine similarity between the embeddings (extracted using CLIP ViT-B/32) of the two sets of images was 0.8190 for Stable Diffusion and 0.7687 for MidJourney. This suggests that both models interpreted the MR-PROMPT and MU-PROMPT similarly, thus creating similar images for both prompts. Conversely, natural images exhibit a much wider range of phenotypic diversity than the images generated with AI models. For example, the images in the `CFD-MR` (Ma, Kantner and Wittenbrink, 2021) showcase greater diversity under the “Mixed Race" category, as illustrated in Figure 7.2. From a quantitative perspective, the mean cosine similarity of embeddings (also extracted using CLIP ViT-B/32) within the `CFD-MR` dataset is 0.7883, whereas the generated images with the MR-PROMPT exhibit larger internal similarity (0.8694 for Stable Diffusion and 0.8265 for Midjourney). Despite the small sample size, our analysis of generative AI outputs provides key insights into how these models replicate and reinforce existing racial stereotypes. More details on dataset composition and comparative analysis can be found in Appendix 7.9.4.2."

Given the apparent limitations of existing frameworks to address race in ML, we present next an alternative approach that aims to more effectively account for racial dynamics and mitigate biases in ML systems.

7.6 MOVING BEYOND RACE LABELS IN ML

In response to the concerns raised in this paper and more generally the constructivist understanding of race (Smedley and Smedley, 2005), we propose that ML research should abandon racial taxonomies where possible and instead fo-

cus on developing alternative methods that address identity-based disparities without reifying race as a biological category.

The challenge to tackle racial discrimination without fixed racial categories has been addressed in multiple disciplines. There is a growing shift towards fine-grained, context-sensitive approaches that challenge rigid racial classifications in a variety of fields, including biology (Yudell et al., 2016), economics (Rose, 2023), public health (Braveman and Parker Dominguez, 2021), social psychology (Cikara, Martinez and Lewis Jr, 2022), and the law (Hu and Kohler-Hausmann, 2024). This perspective aligns with the European approach to combating racism, which largely avoids racial categories in governance, as previously described. However, it is not without challenges. As noted by Braveman and Parker Dominguez, "abandoning the term 'race' has not been accompanied by routine monitoring of health and well-being according to markers of the ethnic groups that are relevant to racism" (Braveman and Parker Dominguez, 2021).[3] This highlights the need for alternative frameworks that capture discrimination while avoiding racial essentialism. What is needed, then, are flexible frameworks with features that can capture discrimination and ensure equal representation based on context-relevant attributes. Following Hu and Kohler-Hausmann, we refer to these as *constitutive features*—attributes that shape the social construct of race— rather than treating race as a fixed variable (Hu and Kohler-Hausmann, 2020). Which features are relevant will depend on the task, the social context, and the geographic setting, leading to a key question: *"Given how a category is constituted, what algorithmic procedures do we consider fair?"* (Hu and Kohler-Hausmann, 2020)

The question above is an interdisciplinary challenge that needs to be addressed in each deployment context. ML research plays an important role in operationalizing these considerations, providing the tools and frameworks to implement fairness interventions that do not rely on racial taxonomies. Below, we outline a research agenda in ML that does not depend on racial classification, focusing on two types of data: (1) tabular and (2) visual data. We conclude with recommendations on how ML can integrate these approaches, emphasizing participatory AI and interdisciplinary collaboration to ensure fairness frameworks reflect real-world disparities without reinforcing essentialist racial categories.

3 It has been found though that perceived race is a better predictor of health disparities than self-reported race (Jones et al., 2008), supporting the focus on racialization.

7.6.1 *Tabular Data*

Algorithmic fairness research in tabular datasets typically evaluates racial discrimination using predefined, fixed race categories. Some studies assume that these categories are unavailable to auditors ("Fairness under Unawareness") (Chen et al., 2019), yet still treat race as an underlying ground truth inferred from proxies. We argue for a shift in perspective: rather than approximating race through proxies, these indicators should be treated as constitutive features of racial discrimination. This approach shares similarities with causal fairness in ML (Kilbertus et al., 2017), but crucially avoids assuming a causal link between "proxies" and race (Hu and Kohler-Hausmann, 2020), recognizing race as a social construct shaped by context-dependent factors.

Recent work in economics has explored approaches along these lines. For instance, discrimination studies in Swiss online recruitment used a composite ethnicity variable—combining language, nationality, and name-based classification—to analyze hiring biases (Hangartner, Kopp and Siegenthaler, 2021)[4]. While this method allows for a more direct examination of discrimination, it remains very limited by its reliance on an overlap of three categorical variables, eventually leading again to a binary comparison.

A more flexible alternative has been proposed by Rose (Rose, 2023), replacing categorical race labels with a *race function* —a context-specific mapping of individual characteristics into a racial space, which assigns a percentage of different perceived races to each attribute combination. This approach is in line with our position as it replaces categorical race variables with constitutive features of perceived race. However, the need to define a function that determines such explicit percentages can also be seen as overly specific. Moreover, the approach was proposed for empirical discrimination research rather than auditing ML models. We propose two possible avenues for integrating these insights into algorithmic fairness.

The first avenue consists of connecting this framework to *individual fairness* (Dwork et al., 2012). Individual fairness enforces consistency by ensuring that similar individuals receive similar predictions, based on a task-specific similarity metric. This metric, which operates in the feature space, is conceptually orthogonal to the race function, which instead defines similarity in the space of perceived race. In analogy with independence-based notions of group fairness (Barocas, Hardt and Narayanan, 2023), fairness could be operationalized as the independence of prediction quality or decisions from the position in this space. Exploring connections between the race function approach and indi-

4 Name-based classification was based on a name-ethnicity recognition algorithm.

vidual fairness could, thus, also be seen as a refinement of group fairness along constructivist lines. We stress that the applicability of the individual fairness approach is not straightforward here; indeed, while it overcomes the problem of rigid taxonomies, its problematic reliance on a specific function has been widely discussed in the Fair-ML literature. Instead, we hope that the race function approach, which was more recently developed in Economics (Rose, 2023) and faces very similar problems, can be refined by drawing on this rich literature.

The second route is through *(multi-)calibration* (Hébert-Johnson et al., 2018), which assesses predictive performance across a vast set of attribute-based groups without relying on predefined race categories. This approach comes with the advantage of not relying on specific functions; however, this also means that it does not straightforwardly allow to incorporate the context of domain-specific discrimination or socialization. As a first step, existing multi-calibration algorithms can be adapted to focus specifically on groups defined by relevant constitutive features, ensuring more granular fairness assessments. Future work should better integrate domain knowledge to refine these subgroup definitions and improve how errors are accounted for within these models. Integrating a constructivist understanding of race (Rose, 2023) into a flexible calibration framework (Höltgen and Williamson, 2023) could help address racialization without enforcing discrete groupings.

7.6.2 *Visual Data*

Computer vision research has increasingly adopted skin tone as a key variable for the study of bias and demographic representation (Bloomberg, 2023; Buolamwini and Gebru, 2018). While this represents progress beyond rigid racial taxonomies, skin tone alone does not fully capture the complexity of racialization and bias in AI (Yucer et al., 2024b).

The idea that racial perception extends beyond skin tone is discussed in the broader literature. Research in cognitive and social psychology shows that racial perception is shaped by multiple phenotypical traits, including skin tone, but also nose shape, eye structure, and lip fullness (Brown Jr, Dane and Durham, 1998; Burgund et al., 2024; Butler, 2011; Roth, 2016; Travers, Fairhurst and Deroy, 2020). These traits interact, such that racial categorization is determined by both visual cues and socio-cognitive processes.

A relevant example is the Face Race Lightness illusion (FRL) (Levin and Banaji, 2006), where prototypical faces —what the original study refers to as *Average White* and *Average Black* faces (Levin, 2000)— are perceived differently in terms of lightness, despite having identical luminance. These stimuli were

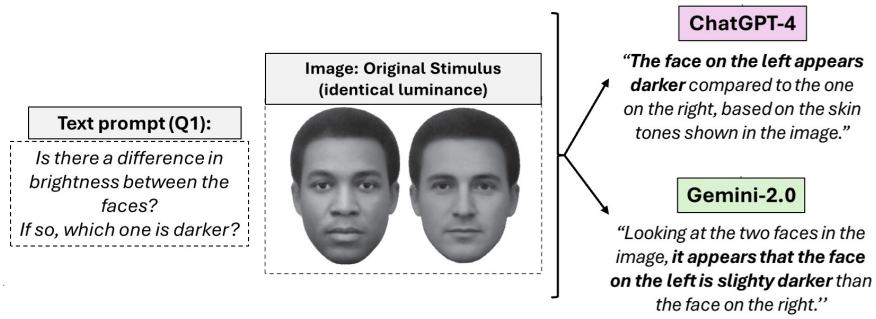

FIGURE 7.3: Face Race Lightness Illusion (Levin and Banaji, 2006) applied to VQA models (ChatGPT-4o, Gemini-2.0-flash-exp)

generated by averaging multiple grayscale male faces through an image morphing process, ensuring that the only differences between them lay in their internal features (eyes, nose, and mouth), while overall brightness and contrast remained controlled.

To illustrate that these psychological findings are also relevant in AI, we tested whether Visual Question Answering (VQA) systems exhibit the same perceptual bias (see Appendix 7.9.5 for the details) as humans. We selected ChatGPT-4 (OpenAI, 2023) and Gemini-2.0 (Google AI, 2024), two widely used VQA models, and presented them with the original FRL stimulus (Levin and Banaji, 2006) testing each model ten times (N=10) to assess response consistency. Both models exhibited the perceptual distortion observed in the FRL study (Figure 7.3). Gemini-2.0 consistently identified one face as darker across all trials, while ChatGPT-4 showed the same pattern in most cases. However, on two occasions (out of 10), ChatGPT-4 did not replicate the illusion and instead reported identical brightness values based on numerical analysis. These results suggest that perceived brightness in these models is influenced by the classification of other facial features, indicating a possible entanglement between phenotypical traits and brightness perception in their latent space. While this example does not imply systematic discrimination, it highlights how facial features can shape skin tone perception, emphasizing the importance of considering such interactions in broader fairness evaluations.

Despite the known influence of phenotypical traits on racial and skin tone perception, the ML community lacks datasets that explicitly annotate diverse phenotypical attributes, and research efforts remain limited and fragmented (Yucer et al., 2024b). For example, the `IBM Diversity in Faces (DiF)` dataset (Merler et al., 2019) introduced detailed facial annotations but was later discontinued after it was revealed that the images had been scraped from online sources without explicit consent, raising serious legal and ethical concerns (Harvey, 2021).

The removal of DiF highlights the difficulties of ethically assembling large-scale facial datasets. Given these difficulties, some researchers have re-annotated existing datasets with phenotypical attributes (Yucer et al., 2020, 2022). In particular, (Yucer et al., 2022) proposes a phenotypic-based framework to replace protected attributes like race. This framework considers traits such as skin tone, eyelid type, nose shape, lip shape, hair color, and hair texture—selected based on social behavior (Feliciano, 2016) and medical studies (Fakhro et al., 2015) (see full categories in Appendix 7.9.6)—allowing them to annotate existing public datasets, such as VGGFace2 (Cao et al., 2018) and RFW (Wang et al., 2019). Their analysis revealed biases in face recognition models, particularly against darker skin tones, wider noses, and monolid eyes, with compounding effects when multiple traits intersect. While this annotation framework has been applied to existing datasets, its use in fairness evaluations remains sparse, and no standardized approach has yet emerged.

To avoid ethical concerns related to collecting phenotypical traits of individuals and to overcome the lack of datasets, some authors have explored the potentials of alternative techniques. GAN-based facial perturbations (Zhang et al., 2022) that manipulate specific features in faces (Georgopoulos et al., 2021; Yucer et al., 2024a, 2020), and feature-masking (*i.e.*, occluding specific facial regions) (Huang et al., 2023; Ozturk, Wu and Bowyer, 2024) have been recently explored to isolate and examine the effects of individual phenotypical traits on model decisions.

While recognizing the potential in these works, we highlight that phenotypical attributes are often presented as proxies for race, which can reinforce an essentialist view of race as a biological category (Crawford and Paglen, 2021; Hanna et al., 2020). Instead, they should be considered important components of racial discrimination.

7.6.3 *Implementation in practice: Context and Participation*

A key challenge in moving away from racial categories is ensuring that fairness interventions remain effective and contextually grounded. A potential solution is the adoption of participatory AI methods, which require direct collaboration with affected communities and domain experts to develop classification criteria and fairness frameworks that reflect specific geographic and socio-political realities. For instance, in Europe, discrimination often targets Romani communities (Fekete, 2014; Trehan and Kóczé, 2009) and operates through regional hierarchies, such as Mediterranean vs. Nordic identities (Levy, 2015; Macharia, 2017), highlighting the need for fairness frameworks tailored to local contexts. How-

ever, meaningful participation requires more than consultation—it must grant real decision-making power to those impacted by AI-driven classification systems (Birhane et al., 2022a; Corbett, Denton and Erete, 2023).

Existing research highlights best practices for fostering meaningful community involvement. Birhane et al. emphasize that participatory approaches should center the knowledge and lived experiences of historically marginalized groups, rather than treating them as late-stage consultants in AI development (Birhane et al., 2022a). Similarly, Delgado et al. document a participatory AI case study in the legal field, where interdisciplinary collaboration between domain experts and computer scientists led to iterative system design improvements, enhancing both effectiveness and fairness (Delgado, Barocas and Levy, 2022). However, these efforts must be carefully implemented to avoid "participation-washing"—where community involvement remains superficial and fails to shift power dynamics—can undermine the intended impact of these initiatives (Sloane et al., 2022).

We propose the following 4 principles, corresponding to the acronym EDTAL, to ensure that participatory AI leads to meaningful fairness interventions: (1) Early and sustained **E**ngagement – community participation should begin at the problem-definition stage, not just during model evaluation; (2) **D**ecision-making power – affected groups should play an active role in shaping AI development, beyond mere consultation; (3) **T**ransparency and **A**ccountability – the rationale and methods for community involvement should be well-documented and publicly accessible to prevent tokenism; and (4) **L**ocalized and context-sensitive approaches – considering regional and cultural differences, as racial biases manifest differently across contexts (Ball, Steffens and Niedlich, 2022; Fekete, 2014).

7.7 ALTERNATIVE VIEWS

The most obvious critique of our proposal is the position that race categories are too important and/or useful to eliminate them. A first argument may especially be raised by those who identify with a marginalized racial group and worry that omitting the categories limits their visibility and their ability to have a voice. A related reasoning was expressed in the hearings following the U.S. decision to allow the indication of multiple races, with groups within the civil rights movement arguing against this plan, based on worries that it may reduce the political effectiveness, despite rejecting the existence of separate biological races (Robbin, 2000; Williams, 2003). Such a position is typically referred to as *strategic essentialism*, a "political strategy whereby differences (within a group) are temporarily downplayed and unity assumed for the sake of achieving polit-

ical goals" (Eide, 2016). Since the 60s, U.S. law now also explicitly provides protection against disparate impact across racial groups in specific contexts, most notably employment (Barocas and Selbst, 2016). Similar arguments to strategic essentialism may be made for the necessity of race categories in ML, for pointing out discrimination in applications like COMPAS.

We think that strategic essentialism is indeed a legitimate and sometimes necessary strategy, as different constituents of race are adequate in different contexts (Roth, 2016). However, the important nuance we would like to stress in this paper is that racial categories in the context of ML should neither be imposed from outsiders in a position of power nor be considered universally applicable. Furthermore, there should be a critical discussion about their construction and a justification for their use (Crisan et al., 2022; Gebru et al., 2021; Mickel, 2024; Mitchell et al., 2019; Okoro et al., 2021). Discrimination studies —such as the analysis of the COMPAS system— often simply interpret race as "perceived race" which does not do justice to complexity of the concept (Hu and Kohler-Hausmann, 2024). Regarding the bearing of disparate impact laws in the U.S. on ML algorithms, there is a concern from the legal perspective "that developers will focus too narrowly on [Statistical Parity Tests], making choices keyed to these metrics, rather than try to understand why disparities are arising and where substantive unfairness may be affecting the selection process" (Raghavan and Kim, 2024). While demographic parity between race categories can already be subject to substantial distribution shifts across different states within the United States (Ding et al., 2021), such failures for fairness criteria to generalize can be expected to be even stronger across geographic regions with substantially different historical biases experienced by racial groups. In sum, our position is that ML researchers and practitioners should be more critical of simplistic race taxonomies and develop methods to analyze discrimination more flexibly, even if there can be situations where their use is warranted.

7.8 CONCLUSION

In this paper, we advocate for the abandonment of race categories in machine learning by default, and call for a fundamental rethinking of how race is conceptualized and operationalized. Building on research from other disciplines, we discuss the ethical, societal, and philosophical challenges of treating race as a categorical "ground truth" both in tabular and visual data. We contrast U.S.-centric existing race taxonomies with practices in Europe and illustrate in the case of mixed-race individuals how current AI practices fail to address the complexities of racial identity while risking to harm marginalized communities.

Further highlighting the problems of reifying and stereotyping race in AI, we call for alternative approaches and suggest research directions for ML. Detecting discrimination is not merely a formal computational task, but also a contextual and normative endeavor (Hu and Kohler-Hausmann, 2024; Selbst et al., 2019; Wachter, Mittelstadt and Russell, 2021). Thus, moving beyond fixed racial categories requires careful scrutiny to ensure that fairness interventions remain effective without reinforcing essentialist assumptions. We call on the ML community to critically engage with these challenges and develop the awareness and tools necessary for a more nuanced and equitable approach.

Author's positionality

We recognize the importance of reflecting on our own identities and experiences in shaping this work. Our team consists of four authors—three identifying as female and one as male—with three different European nationalities. Collectively, we have lived, studied, and worked across various international contexts, both inside and outside Europe, which informs our understanding of global perspectives on race and technology. The first author, being of European-African heritage, brings personal insight into the lived experiences of navigating racial identity, directly informing the critical analysis presented in this paper. Our combined expertise spans over 30 years of research in artificial intelligence, computer vision, human-computer interaction, AI for social good, algorithmic fairness, and philosophy of science, grounding our exploration of race taxonomies in AI within both technical and ethical frameworks. We acknowledge the limitations of our perspectives and remain committed to ongoing reflection and engagement with the communities most impacted by these technologies.

7.9 APPENDIX

7.9.1 *Race-Related Datasets used in ICML and CVPR Papers (2023-2024)*

Table of the 28 accepted papers that incorporate race categories in their research from ICML (featuring 'fair' in the title) and CVPR (featuring 'fair' or 'bias' in the title), along with the datasets utilized.

TABLE 7.1: List of race-related datasets used in fairness research, indicating the papers and their respective conferences (CVPR = * and ICML = *). (2023-24)

Race Dataset	**Papers Using It *= ICML, *=CVPR (2023-24)**
FairFace (Karkkainen and Joo, 2021)	[Chen et al., 2024, Qraitem, Saenko and Plummer, 2023, Garcia et al., 2023, D'Incà et al., 2024]*
UTKFace (Zhang, Song and Qi, 2017)	[Qraitem, Saenko and Plummer, 2023, Park and Byun, 2024]*; [Nelaturu et al., 2024]*:
COMPAS (ProPublica, 2016)	[Soen, Husain and Nock, 2023, Memarrast, Vu and Ziebart, 2023, Singh et al., 2023, Zhu et al., 2023, Peltonen et al., 2023, Roh et al., 2023, Kim and Zubizarreta, 2023, Becker, Dumitrasc and Broelemann, 2024, Tifrea et al., 2024]*
Census (Adult / Folktables) (Ding et al., 2021)	[Xian, Yin and Zhao, 2023, Khalili, Zhang and Abroshan, 2023, Jovanović et al., 2023, Singh et al., 2023, Okati, Tsirtsis and Rodriguez, 2023, Knittel et al., 2023, Keswani, Mehrotra and Celis, 2024]*
MINNEAPOLIS [5]	[Soen, Husain and Nock, 2023] *
FASSEG (Khan et al., 2015)	[Zhang, Roth and Zhang, 2024]*
HSLS (Jeong et al., 2022)	[Tifrea et al., 2024]*
ENEM (Alghamdi et al., 2022)	[Tifrea et al., 2024]*
Law School (Wightman, 1998)	[Peltonen et al., 2023, Xu and Strohmer, 2023, Xian et al., 2024] *
Communities & Crime (Wightman, 1998)	[Xian, Yin and Zhao, 2023, Singh et al., 2023, Xu and Strohmer, 2023, Giuliani, Misino and Lombardi, 2023, Xian et al., 2024, Tifrea et al., 2024] *
Toxic Comments (Borkan et al., 2019)	[Xu et al., 2024] *

7.9.2 *U.S. census categories*

Since 1997, race/ethnicity data was collected in the U.S. census as depicted in Figure 7.4 (US Office of Management and Budget, 1997).

In 2024, the Statistical Policy Directive No. 15 on Race and Ethnicity Data Standards was amended for the first time since 1997. Among other changes, a new category “Middle Eastern or North African" was added, which was previously subsumed under “White".

Are you Hispanic or Latino?

☐ No, not Hispanic or Latino

☐ Yes, Hispanic or Latino

What is your race? ***Select one or more.***

☐ American Indian or Alaska Native

☐ Asian

☐ Black or African American

☐ Native Hawaiian or Other Pacific Islander

☐ White

FIGURE 7.4: U.S. census questions for race and ethnicity after 1997.

7.9.3 *Additional information on datasets*

`ColorFERET` (Phillips et al., 2000) use *White, Asian, Black, Others*

`MORPH` (Ricanek and Tesafaye, 2006) use *Caucasian, Hispanic, Asian, African American*

`UTKFace` (Zhang, Song and Qi, 2017) use *Asian, Black, Indian, White, Others*

`FairFace` (Karkkainen and Joo, 2021) use *Black, East Asian, Indian, Latino, Middle Eastern, Southeast Asian, and Western*; the authors follow the U.S. census standard to subdivide White into Western and Middle-Eastern, but see Asian as subdivided into East Asian and Southeast Asian, with Indian as an independent race, contrary to the U.S. census.

`Adult / folktables` (Becker and Kohavi, 1996; Ding et al., 2021) is directly derived from census data (with mixed-race coded as single category); the old `Adult` data only had *White, Asian-Pacific-Islander, American-Indian-Eskimo, Other, Black*. In the new `folktables` version, the full category names are *White, Black*

TABLE 7.2: Race categories utilized in popular Machine Learning datasets. More information in Appendix 7.9.3. The second column indicates the number of categories, the third column indicates whether categories are a subset of the U.S. Census categories.

Dataset	#	U.S.	Note
ColorFERET	4	√	
MORPH	4	√	
UTK Face	5	√	*Indian* separated from *Asian*
FairFace	7	√	*White* & *Asian* subdivided, *Indian* separated
Adult / folktables	5/9	√	
COMPAS	6	√	
Communities & Crime	4	√	
LSAC Law School	8	×	plus *Mexican American* & *Puerto Rican*

or African American, American Indian, Alaska Native, American Indian or Alaska Native[6]*, Asian, Native Hawaiian and Other Pacific Islander, Other Race, Two or More Races*

`COMPAS` (ProPublica, 2016) uses *African-American, Asian, Caucasian, Hispanic, Native American, Other* based on police/judicial data

`Communities & Crime` (Redmond and Baveja, 2002) uses U.S. census data for the percentages of race categories per community, using only *black, white, asian, hispanic*

`LSAC Law School` (Wightman, 1998) divides by "Ethnic Group" categories and uses *American Indian, Asian American, Black, Mexican American, Puerto Rican, Hispanic, White, Other*

For all the tabular datasets, it is also common to binarize into white/non-white or black/non-black (in the case of `Communities & Crime`) or to use any other subset. This is despite the Statistical Policy Directive No. 15 from 1997 declaring that "[t]he term 'nonwhite' is not acceptable for use in the presentation of Federal Government data."

6 Full: American Indian and Alaska Native tribes specified, or American Indian or Alaska Native, not specified and no other races

7.9.4 *Mixed-Race Identity*

7.9.4.1 *Definition of "mulatto"*

The inclusion of the term *"mulatto"* in this study was intentional and motivated by its definition in the Oxford English Dictionary (Oxford English Dictionary, 2025b).

> "mulatto, n. & adj. A person with one white and one black parent. Frequently more generally: a person of mixed white and black ancestry. Cf. metis, n. A.1, quadroon, n. Now chiefly considered offensive."

This term, while offensive in contemporary English, reflects a historical framing of mixed-race identities within a narrow Black/White binary. We hypothesize that Generative AI systems, trained on datasets infused with historical and cultural biases, may replicate or reinforce such reductive categorizations.

7.9.4.2 *Additional information regarding the Midjourney and Stable-Diffusion 3.5 Large experiments*

To investigate the representation of mixed-race individuals, we conducted an experiment using Stability AI's Stable Diffusion 3.5 Large (Huggingface, 2024), Midjourney (MidJourney, 2024), and OpenAI's DALL·E 3 (OpenAI, 2024). The experiment focused on two specific prompts: *"a mixed-race person"* (MR-prompt) and *"a mulatto person"* (Mu-prompt). For each prompt, we generated 30 images per model, resulting in two distinct datasets for comparison.

As real images, we used the 88 images categorized as "Mixed-Race" in the Chicago Real Face Database (Ma, Kantner and Wittenbrink, 2021). To ensure consistency, all real and generated images were cropped before feature extraction, focusing on the facial area

TABLE 7.3: Cosine similarity distribution between images generated for "a mixed-race person" (MR-PROMPT) and "a mulatto person" (MU-PROMPT) in Stable Diffusion and MidJourney.

Model	Mean	Median	75th Percentile	90th Percentile
Stable Diffusion	0.8190	0.8253	0.8618	0.8849
MidJourney	0.7687	0.7732	0.8228	0.8975

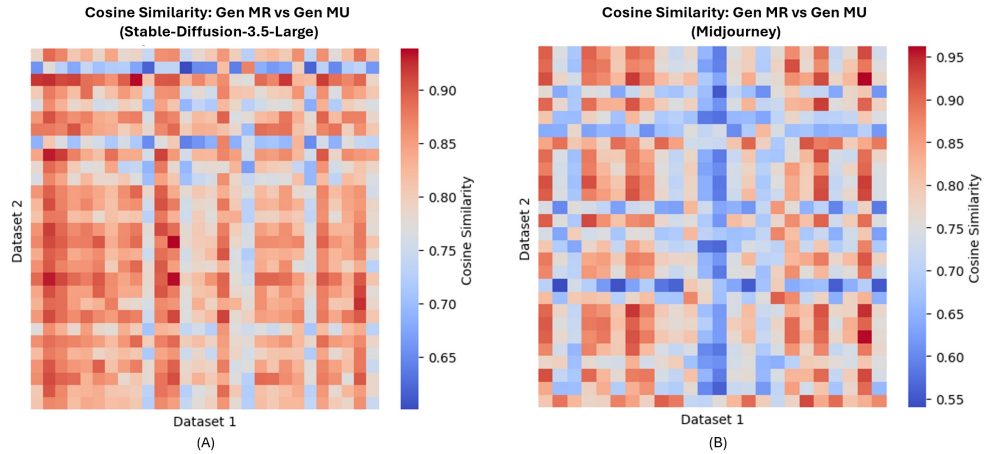

FIGURE 7.5: Heatmap of cosine similarity between images generated using the prompts *"mixed-race person"* (MR-PROMPT) and *"mulatto person"* (MU-PROMPT). On the left (A), results from Stable Diffusion 3.5-Large; on the right (B), results from MidJourney.

TABLE 7.4: Cosine similarity statistics for real and AI-generated mixed-race images. The generated images (SD = Stable Diffusion-3.5-Large and MDJ = MidJourney) exhibit higher internal similarity compared to real-world images from CFD-MR, indicating a lower degree of phenotypic diversity.

Dataset	**Mean**	**Median**	**75th Perc**	**90th Perc**
Real Mixed-Race (CFD-MR)	0.7883	0.7911	0.8489	0.8831
Generated Mixed-Race (SD)	0.8694	0.8770	0.9050	0.9272
Generated Mixed-Race (MDJ)	0.8265	0.8388	0.8933	0.9209

7.9.5 *Additional information on testing the Face Race Lightness Illusion in VQA Models*

7.9.5.1 *Original test*

(Levin and Banaji, 2006) tested this effect by using grayscale images of faces, carefully controlled for luminance, and asked human participants to adjust the brightness of one face to match another. Their results showed that participants consistently perceived Black faces as darker than White faces, even when their objective brightness was the same. This suggests that phenotypical traits can override purely physical visual cues in human perception.

7.9.5.2 *Adaptation*

Inspired by these findings, we investigated whether Visual Question Answering (VQA) models exhibit similar perceptual distortions when processing faces with different phenotypical traits but identical luminance. While AI models cannot adjust brightness levels as human participants did in the original study, they can evaluate and compare brightness differences based on learned visual representations. Thus, our goal was to examine whether VQA models display a bias in perceived brightness similar to that observed in humans. (models: OpenAI's ChatGPT-4 (OpenAI, 2023), Google's Gemini-2.0-flash-exp (Google AI, 2024)). We employed the same stimulus used in the original FRL experiment (Levin and Banaji, 2006). Specifically, we presented the models with the validated pair of average men faces from Levin and Banaji (2006) 7.3.

Our choice to use this specific stimulus ensures a direct replication of the original study in an AI context, avoiding the need for additional psychological validation that new stimuli would require.

We asked each model the following question: **Q1:***"Is there a difference in brightness between the faces? If so, which one is darker?"* . This query was repeated ten times **(N=10)** per model to assess response consistency.

- Gemini-2.0 consistently identified the left face as darker in all ten trials, fully replicating the perceptual distortion observed in human participants.
- ChatGPT-4 showed a similar pattern in eight out of ten trials but deviated twice, providing an analytical response: *"The average brightness of the left face is approximately 196.32, while the average brightness of the right face is about 196.32. Both faces appear to have identical brightness levels."* This suggests that ChatGPT-4 sometimes relied on numerical analysis, which overrode the perceptual distortion.

To further investigate the reasoning behind the models' responses, we asked: **Q2:***"Why do you think the left face appears darker?"* . Both models attributed the effect to differences in luminance and contrast, but with slight differences in emphasis. Gemini 2.0 provided a more structured explanation, explicitly citing melanin concentration and grayscale rendering as primary factors, while ChatGPT-4 offered a more general response, framing the effect in terms of brightness differences and perceptual biases.

> "The left face appears darker primarily due to the amount of pigment in the skin. This is usually due to a higher concentration of melanin, which is a natural pigment responsible for skin, hair, and eye color. In the case of these images, the left face likely belongs to someone with darker skin, leading to a darker appearance in the grayscale rendering of the image.
>
> Here are some supporting points:
>
> - Melanin Content: The amount of melanin in an individual's skin is the primary factor determining skin darkness. People with darker skin tones have higher levels of melanin.
> - Grayscale Imaging: When color images are converted to grayscale, the brightness level of each pixel corresponds to the intensity of the color. Higher levels of pigment (like melanin) absorb more light, resulting in a darker pixel in the grayscale image.
> - Visual Perception: Our eyes and brains interpret darker areas as having less light reflecting off them, and consequently, we perceive them as darker.
>
> Therefore, the darker appearance of the left face is primarily due to the higher concentration of melanin in the individual's skin, which translates to a darker shade in the grayscale image." (Example:*Gemini 2.0*)

> "The left face appears darker due to differences in luminance and contrast. The grayscale shading used in the image likely has lower brightness levels on the left face compared to the right. Additionally, our perception might be influenced by cognitive biases and expectations related to facial features and shading." (Example:*ChatGPT-4*)

7.9.6 *Phenotypic-based framework*

Phenotypical attributes used by Yucer et al. (2022) to label VGGFace2 and RFW datasets inspired by the relevant categories of Feliciano (2016).

TABLE 7.5: List of phenotypical attributes and their respective categories used by Yucer et al. (2022)

Attribute	**Categories**
Skin Type	Type 1 / 2 / 3 / 4 / 5 / 6 (Fitzpatrick Skin Types (Sachdeva, 2009))
Eyelid Type	Monolid / Other
Nose Shape	Wide / Narrow
Lip Shape	Full / Small
Hair Type	Straight / Wavy / Curly / Bald
Hair Colour	Red / Blonde / Brown / Black / Grey

ACKNOWLEDGEMENTS

We thank all those who shared their time and thoughts with us during the early stages of this work, especially members of the mixed-race community, whose perspectives deepened our understanding of racial categorization and the nuanced experiences of in-betweenness.

MD acknowledges support from the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche), and funding from the FNRS (National Fund for Scientific Research) for her visiting research at the ELLIS Alicante Foundation.

BH is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. BH also acknowledges support by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and travel support from ELIAS (GA no 101120237).

PR and NO are supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación,

Industria, Comercio y Turismo, Dirección General de Innovación). PR is also supported by a grant by Fundación Banc Sabadell.

Part IV

EXTENDING THE SCOPE OF PREDICTIONS

8

THE ACCOUNTABILITY GAP IN CAUSAL MODELLING FOR AUTOMATED DECISION SYSTEMS

In general, we consider it a good principle to explain the phenomena by the simplest hypotheses possible, in so far as there is nothing in the observations to provide a significant objection to such a procedure.

— Ptolemy, *Almagest*

AUTHOR CONTRIBUTIONS

The idea for the project, conceptualisation, theoretical results, and writing are due to Benedikt Höltgen. Robert Williamson supervised the project and gave valuable feedback.

ABSTRACT

Automated decision systems increasingly rely on causal modelling to inform high-stakes decisions in domains such as health and social policy. Despite the goal of informing decisions about a concrete population, the underlying inductive assumptions take the form of abstract sampling and independence assumptions, without modelling the target population directly. This reliance on abstract assumptions makes it difficult to hold decision makers accountable for their modelling choices. In this paper, we develop a variant of the potential outcomes framework due to Neyman and Rubin in which we construe causal inference as treatment-wise prediction on a target population. All assumptions are formulated in terms of observable quantities and are testable in retrospect; this makes it possible not only to evaluate predictions, but also to diagnose sources of error when predictions fail. We show that the proposed framework captures established causal inference methodology. By shifting the focus from identifying parameters of abstract distributions to predicting outcomes for concrete popula-

tions, the framework supports more transparent evaluation and accountability for automated decision systems based on causal models.

8.1 INTRODUCTION

For many problems in the social and health sciences, it is important to analyse and predict the efficacy of treatments or policies. This requires causal modelling, as observational outcome distributions may not reflect outcome distributions under active treatment policies. It is also increasingly recognised in the literature on algorithmic or automated decision systems (ADSs) that purely predictive systems do not model the effects of the very decisions they inform (Kern et al., 2025; Liu et al., 2025; Wang et al., 2024). Across many fields, Rubin causal models (RCMs), due to Neyman (Neyman, 1923/1990) and Rubin (1974), are the preferred causal framework for informing specific policies or interventions, due to the focus on specific outcomes of interest and the aptitude to accommodate individual problem settings (Imbens, 2020; Markus, 2021)—*vis-à-vis* the conventional econometric approach (Heckman and Pinto, 2024) and structural causal models (Pearl, 2009), which are sometimes seen as having the edge in more abstract theory building as they model unobservables more explicitly. In line with this, the adoption of RCMs is commonly also advocated for ADSs that are used for the allocation of interventions such as job trainings (Liu et al., 2025; Zezulka and Genin, 2024).

Despite their focus on outcomes of interest and the aspiration to 'carefully describe their sample of trials and the ways in which they may differ from those in the target population' (Rubin, 1974, p. 698), work using RCMs often neglects the analysis of external validity, that is, generalisation to a target population (as lamented across fields, see Section 8.2.3). As a result of this, it is difficult and rare to make testable claims about the adequacy of causal ADSs. We argue that this is in part due to the framework itself, as it—like its alternatives— does not provide a straightforward way to model target populations. Instead, the focus of the framework is the 'identification' of parameters in some abstract probability distribution—'as if a parameter, once well established, can be expected to be invariant across settings' (Deaton and Cartwright, 2018, p. 10). A further issue with this abstract framing is that it is very difficult, if possible at all, to formulate concrete and testable assumptions that enable the accurate prediction of the effects of policies on target populations as well as critical assessment of modelling assumptions and choices. This issue is aggravated by the fact that the evaluation of claims about individual causal effect is even more challenging than that of probabilistic predictions (Vafa, 2024).

This poses a problem for the accountability for ADSs that are informed by causal modelling. Accountability is most commonly understood as a relation between an actor and a forum (Bovens, 2010)—this may be regulatory agencies or the public (Metcalf et al., 2023). For algorithmic accountability specifically, it means that 'multiple actors (e.g. decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct' (Wieringa, 2020, p. 10). These actors should communicate assumptions, evaluations, and the scope of their models, e.g. in standardised formats like model cards (Mitchell et al., 2019), to assess 'improper assumptions in design and use and other human and organisational factors' (Cobbe, Lee and Singh, 2021, p. 600). It has not yet been explored what this means for causal models with their aforementioned difficulties of evaluate predictions and assumptions, despite their growing adoption.

In this paper, we draw attention to this gap and suggest a way to overcome it. More concretely, we suggest a variation of RCMs that directly models both observed and target populations and their relationship. It shifts the focus from the identification of true parameters to the prediction of future outcomes based on (retrospectively) testable assumptions. Rather than ignoring existing causal inference methodology, we show that the new framework can capture established estimators and provide a complementary perspective on them. We, thus, provide an 'intermediary' framework that establishes links between high-level intuitions, as formalised in RCMs, on the one hand and directly testable assumptions about concrete populations on the other. Hence, this variant augments the strengths of RCMs for evidence-based policy making by focussing on predictions for concrete target populations grounded in testable assumptions.

The structure of this work is as follows. In Section 8.2, we recapitulate RCMs and discuss calls across fields to direct more attention to problems with external validity and to model the target population more directly. In Section 8.3, we introduce the new framework in the context of simple estimators; a broader survey of existing estimators and how they fit into the new framework can be found in Appendix 8.7.1. In Section 8.4, we compare the new with the conventional framework, by drawing formal connections and discussing differences in practice. While most of the paper works with *average* treatment effects and potential outcomes for ease of presentation, Section 8.5 explains how this generalises to conditional treatment rules. Section 8.6 concludes.

8.2 BACKGROUND: RUBIN CAUSAL MODELS (RCMS)

In this section, we first introduce the standard formalism of RCMs, before critically discussing the assumption of an underlying probability distribution as well as the problem of explicitly modelling outcomes on a target population.

8.2.1 *The standard framework*

The main components of RCMs are the following random variables (RVs): T_i is the decision variable indicating whether person i is treated and takes values in $\{0, 1\}$, which represent control and treatment.[1] Y_{1i} and Y_{0i} denote the outcome for i upon receiving treatment and control, respectively. Based on this, we can define the actual outcome

$$Y_i := T_i \cdot Y_{1i} + (1 - T_i) \cdot Y_{0i} = Y_{0i} + T_i(Y_{1i} - Y_{0i}). \tag{8.1}$$

Consider the common example of labour market programme evaluation: T_i indicates whether someone gets offered job training and Y_i indicates whether they have a job after a fixed time, say, one year. It is, thus, assumed that for every person, both Y_{0i} and Y_{1i} are well-defined, i.e., whether they find a job *if* they don't get offered job training and whether they find a job *if* they are assigned job training, respectively. Of course, we can, in principle, only observe the value of one of the two variables for each individual.

In many settings (some of which we will consider in this paper), we also have informative covariates X_i. In the context of job trainings, these may be attributes such as age, gender, education, and employment history. A foundational assumption behind RCMs is that there is a joint distribution P over all variables, i.e.

$$Y_{1i}, Y_{0i}, T_i, X_i \sim P. \tag{8.2}$$

We are then usually interested in the average treatment effect (ATE) $\mathbb{E}_P[Y_{1i} - Y_{0i}]$. The ATE is typically seen as the expectation over the individual treatment effect (ITE) $Y_{1i} - Y_{0i}$. The fact that we can only ever measure one of them for each i has been dubbed the 'fundamental problem of causal inference'.

In line with many textbooks, we showcase RCMs in the context of Randomised Controlled Trials (RCTs). This represents the 'experimental ideal' in the sense that it assumes we have access to data from a randomised experiment.

1 We only consider the binary treatment case here as this makes the presentation of RCMs simpler.

We can then assume that the potential outcomes are independent of the treatment decision

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp T_i. \tag{8.3}$$

This means we don't need covariates X_i here, as the ATE can be expressed as

$$\mathbb{E}_P[Y_{1i} - Y_{0i}] = \mathbb{E}_P[Y_i|T_i = 1] - \mathbb{E}_P[Y_i|T_i = 0], \tag{8.4}$$

and both terms on the RHS can directly be estimated from our data: Assuming that the data are sampled from the distribution P, the law of large numbers says that the empirical estimate of (8.4) converges to the ATE almost surely.

RCTs are often not available and many methods have been devised to allow causal inference when random assigment and thus the conditional independence (8.3) is violated. Most of these methods rely on the assumption of *unconfoundedness* (sometimes also called conditional independence, ignorability, or selection-on-observables), given by

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp T_i \mid X_i. \tag{8.5}$$

This means that treatment assignment may have been based on X_i but not on any other variables that could give information about the outcome. In other words, the treatment group should be comparable to the control group if we take the covariates into account. For example, in labour market programmes, information about potential participants is taken into account for deciding who gets offered job training. The unconfoundedness assumption is then assumed to be satisfied if the decision was only based on the covariates X_i—hence the name 'selection on observables'.

8.2.2 *The distribution*

As described above, RCMs are formulated in terms of joint distributions that represent the ground truth and are used to derive theoretical guarantees. But how should these distributions be understood? Should we take them at face value and believe that they are supposed to correctly describe some data-generating process? Or are they just models that have a more indirect relationship with what we believe to be true about the world? We now discuss these two options in turn.

The first option is that descriptions in terms of generative distributions can be true or false. Much of the literature suggests this reading. This would mean e.g. that there is a true (marginal) distribution of covariates from which people are sampled. And this seems to hold then for any selection of covariates/attributes

that we can come up with. In principle, the number of possible choices of covariates is almost unlimited, although we often just take the attributes that we can most easily measure. The invoked distributions are commonly not thought to pertain to a specific time, though sometimes to a specific location. But the relationships between e.g. unemployment, age, and education clearly change over time—if they can be said to be stable at any given point in the first place. Every case of a person finding or not finding a job is highly individual and it seems rather strong that they just follow a general law plus some noise—a notion we know from physics. Nancy Cartwright (1999) makes the point that while a lot of work in physics has focused on finding latent quantities that do have stable relationships (like forces in Newton's and Hooke's laws), social sciences like economics typically consider more easily measurable quantities that are particularly salient.[2] And 'to suppose that there really is some probability measure over [such quantities], you need a lot of good arguments' (p. 325).[3]

The other option is, then, to say that aspects like sampling from a joint distribution are mere models that are always wrong (if they have truth conditions at all). On this view, assumed joint distributions are idealisations and aim to capture patterns that we can observe between individual events, but they cannot be literally true. If this is to be the interpretation, what does e.g. the dependence on unconfoundedness mean if it can never be true? If generative distributions are useful fictions, under which conditions are they useful? The idea of data-generating distributions goes back to Fisher, who introduced the notion of hypothetical infinite populations (Fisher, 1922); he saw them as 'products of the statistician's imagination' (Fisher, 1955, p. 71). Already in Fisher's days, 'the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money' (Fisher, 1955, p. 70).[4] Furthermore, he argued that 'the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful, and the identification of the two sorts of operation is decidedly misleading'. In line with this, we contend that the lack of reliable scientific foundation in most ADSs should make us wary of the reliance on abstract distributions and untestable independence assumptions.

2 This applies particularly to RCMs which, in contrast to other causal modelling approaches (Heckman and Pinto, 2024) do not model unobservables.

3 Kolmogorov also argued that 'generally speaking there is no ground to believe that a random phenomenon should possess any definite probability' (Kolmogorov, 1983).

4 This connection between algorithmic systems and saving money is also present in modern discussions (Allhutter et al., 2020).

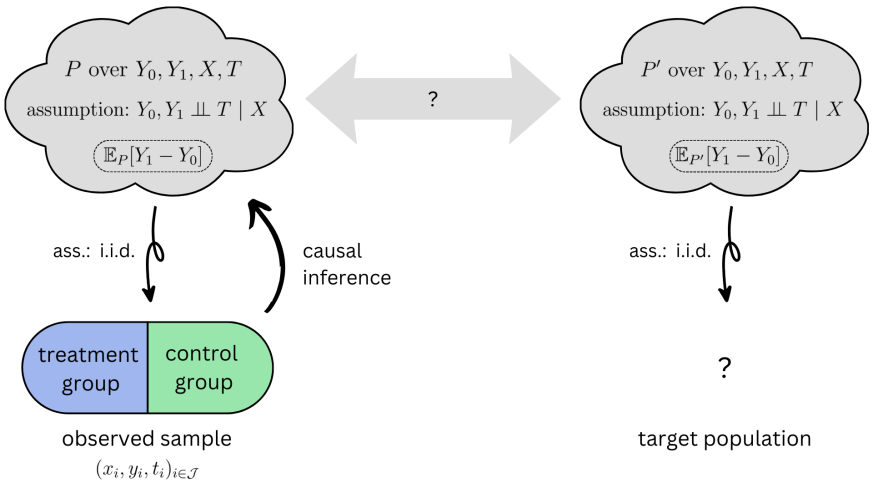

FIGURE 8.1: A schematic drawing of the RCM approach.

While a standard and seemingly innocuous aspect, these distributions actually do much of the heavy lifting. They provide the connections between both the quantities of interest and between different populations, via sampling assumptions. Getting the former right requires internal validity while getting the latter right requires external validity; in RCMs, these two aims are typically considered in separation. Internal validity is concerned with the assumptions discussed above, especially with whether unconfoundedness holds in the assumed 'generative process' behind observational data. This is supposed to ensure that estimations of quantities like the ATE are correct for the observed sample. External validity, in contrast, concerns the question whether these estimations also hold for other data, for example, for populations on which we want to make treatment decisions in the future. It is usually assumed that the distribution reflected in our historical data is the same as or similar to the distribution from which future data is 'sampled', which allows us to predict outcomes or treatment effects in some population of interest. Figure 8.1 provides a visual sketch of the RCM picture, where the left part covers internal validity and changes for the target populations are formulated in terms of distributional changes. The notion of distributions is analogous to that of hypothetical populations in statistics, although nowadays 'the "hypothetical" part of this notion—which is crucial to the concept—gets dropped' (Kass, 2011, p. 3).

8.2.3 *The target population*

As noted by Rubin in the paper introducing the framework, 'for almost any study to be of interest, the results must be generalizable to a population of trials'

(Rubin, 1974, p. 698). However, it has often been observed that '[s]ocial scientists frequently invoke external validity as an ideal, but they rarely attempt to make rigorous, credible external validity inferences.' (Findley, Kikuta and Denly, 2021, p. 365). It is striking how pervasive such statements are across various fields in which RCMs are used. In economics, '[r]esearchers tend to focus primarily on threats to internal validity' (Bo and Galiani, 2021, p. 274), whereas external validity is virtually not discussed in standard textbooks like Angrist and Pischke (2010) (or in general textbooks like Imbens and Rubin (2015)). Similar statements have been made about epidemiology (Lesko et al., 2020, p. 2), public health (Steckler and McLeroy, 2008, p. 9), behavioural medicine (Glasgow et al., 2006, p. 106), or political science (Egami and Hartman, 2023, p. 1070).

One reason is that in Rubin's models, it is difficult to follow Rubin's call for 'investigators [to] carefully describe their sample of trials and the ways in which they may differ from those in the target population' (Rubin, 1974, p. 698), given the difficulty to connect such a target population with the observed data in the formalism. The easiest setting would be to assume that the target population is sampled i.i.d. from the same distribution as the observed population. In this case, one can draw connections between them through the conjured distribution, using finite sample theory to estimate quantities of interest and possible errors or confidence intervals. This, however, is a very strong assumption. To say instead that the target population has a somewhat different generative process would mean to conjure a second abstract distribution that is somehow related to the first, from which the target population is then sampled. Making the relations between them explicit is difficult since none of the relations observed sample – first distribution – second distribution – target population is observable (Figure 8.1) and no assumptions about them can be tested.

Given the observed neglect of external validity, some researchers have recently started to argue for more explicit modelling of the target population. For example, Rudolph et al. (2023, p. 4) argue that 'all statements regarding representativeness should make clear the way in which the study results generalise, the target population the results are being generalised to, and the assumptions that must hold for that generalisation to be scientifically or statistically justifiable'. Similar points have also been made by Westreich et al. (2019) and Fox et al. (2022).

Munger (2023) also criticises 'contemporary external validity literature', arguing that 'prediction that does not acknowledge the scientist's temporal position [..] is merely counterfactual history' (p. 2) and that 'a necessary first step is to *actually* make predictions about the future' (p. 7). In the following, we propose an amendment to RCMs which allows developers to directly model

the target population and to connect it to observed data without any detour through abstract distributions. This incentivises the engagement with concrete conditions that allow generalization and makes assumptions explicit and, retrospectively, testable. Still, the proximity and direct relation to RCMs makes it possible to keep trained intuitions and methodologies.

8.3 NEW FRAMEWORK

8.3.1 *Setup and notation*

Reflecting the aforementioned problems, we now introduce an amendment to the conventional framework; the notation is similar to that of Manski (2004). By $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{T}$ we denote the sets of possible covariates, outcomes (in $\mathbb{R}$), and treatments, respectively. We only consider binary treatments $\mathcal{T} = \{0, 1\}$ in the main text for easier comparison with RCMs, although our framework applies straightforwardly to multiple treatments. We assume that we have datapoints $(x_i, y_i, t_i)_{i \in \mathcal{J}}$ where $\mathcal{J}$ serves as the index set of our training data. That is, for each unit $i \in \mathcal{J}$ with covariates $x_i \in \mathcal{X}$, we have observed outcome $y_i \in \mathcal{Y}$ under treatment $t_i \in \mathcal{T}$.[5]

We then consider a target population $\mathcal{I}$ with an unknown **outcome function** (also called 'response function' (Manski, 2004))

$$y : \mathcal{I} \times \mathcal{T} \to \mathcal{Y} \tag{8.6}$$

such that $y(i, t)$ denotes the (potential) outcome when treatment $t \in \mathcal{T}$ is assigned to individual $i \in \mathcal{I}$.[6] Many methods for non-experimental settings require covariates; we use a **representation function**

$$x : \mathcal{I} \to \mathcal{X} \tag{8.7}$$

to denote covariates of target units by $x(i), i \in \mathcal{I}$. Using a function for these quantities emphasises that not just the outcomes but also the covariates of the target population may not yet be well-defined at the time models are trained and policies for $\mathcal{I}$ are chosen, and that representing an individual i through some covariates x is itself a deliberate action.

5 This means that by default, there is no notation for counterfactual outcomes of observed datapoints.

6 The common SUTVA assumption precluding interactions between treatment assignments to different individuals is encoded in the fact that the outcome only depends on the individual's treatment. One could allow such interactions by taking as inputs i and a treatment vector of length $|\mathcal{I}|$.

As mentioned above, the most common quantity of interest is the average treatment effect (ATE); the finite-population version for the target population in our setting would be

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,1) - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,0). \tag{8.8}$$

This theoretical quantity is the difference between the average outcomes of assigning either treatment or control to everyone, the **average potential outcome (APO)**

$$\mu(\mathcal{I},t) := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t). \tag{8.9}$$

Our approach frames assumptions and results in terms of observable quantities that can be directly compared. To this end, we introduce a shorthand notation for approximate equality:

Definition 8.1. *For $\epsilon > 0$, we say that two values $r, s \in \mathbb{R}$ are ϵ-similar if $|r-s| < \epsilon$. We write this as*

$$r \approx_\epsilon s. \tag{8.10}$$

8.3.2 *Experimental setting: RCTs*

We now turn to assumptions that allow us to make predictions about the target population based on observed data. In the comparatively simple case of RCTs, we do not need covariates for predicting the APO. Here, individuals are randomly assigned into treatment and control group. In the earlier example, this could mean that a lottery is used to decide who is offered job training in a population of unemployed people. We denote treatment ($t = 1$) and control ($t = 0$) group in the observed data $\mathcal{J}$ by $\mathcal{J}_1$ and $\mathcal{J}_0$, that is,

$$\mathcal{J}_t := \{i \in \mathcal{J} : t_i = t\}, \quad t \in \mathcal{T} = \{0,1\}. \tag{8.11}$$

The random assignment in RCTs can be used to justify the assumption that the average outcome in $\mathcal{J}_t$ approximates our quantity of interest (8.9), that is, the APO in the target population $\mathcal{I}$ if we assign treatment t to everyone:

$$\mu(\mathcal{I},t) := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t) \approx_\epsilon \frac{1}{|\mathcal{J}_t|}\sum_{i\in\mathcal{J}_t} y_i. \tag{8.12}$$

For this, we do not need to invoke any true distribution; it is enough to assume that the partition into observed and target samples as well as the partition into control and treatment groups can be considered random. That is, if the partition

into $\mathcal{I}$ and $\mathcal{J}$ is random and the partition of $\mathcal{J}$ into $\mathcal{J}_0$ and $\mathcal{J}_1$ is random, then $\mathcal{J}_t$ under treatment t is representative[7] of $\mathcal{I}$ under treatment t: This means we treat them as being drawn from the same urn, similar to Neyman's original work.[8] This justifies (8.12) because (for large enough data sets) the vast majority of possible partitions will lead to roughly equal averages in both parts. This can be shown with a Hoeffding inequality for finite samples (as in Proposition 1.2 of (Bardenet and Maillard, 2015)), giving statements of the form 'for 95% of partitions, the difference between means is below 0.05'.

Such a measure on the number of admissible partitions can be made into a probabilistic statement ('with 95% probability...') if we additionally make the assumption that all partitions are equally likely—which is made in RCMs in the form of the i.i.d. assumption and in Bayesian frameworks such as (Dawid, 2021) in the form of exchangeability. In a finite population framework, this can, however, be neatly generalised to allow biased sampling schemes, as shown in the literature on non-probability sampling (Meng, 2018, 2022); we elaborate on the connection to this literature in Appendix 8.7.4. Our approach thus accentuates this equi-probability assumption and allows the modeller to easily incorporate deviations, in the form of correlations between outcome and group membership. In our running example of job trainings, this could take the form of incorporating e.g. fluctuations in general unemployment if the data collection happened a few years earlier.

8.3.3 *Predictions and assumptions*

Causal inference becomes more complicated outside controlled experiments and most literature grapples with observational or quasi-experimental settings. For example, data about the efficacy of job trainings typically does not come from RCTs. In such settings, we incorporate additional information in the form of covariates $x \in \mathcal{X}$. As we will demonstrate, these covariates are used not to analyse the difference between treatment and control groups, but between the treatment/control parts of the observed sample on the one hand and the full target sample on the other.

The last important concept in our framework is a predictor

$$p : \mathcal{X} \times \mathcal{T} \to \mathbb{R}. \tag{8.13}$$

7 Rudolph et al. (2023) argue for such a notion of representativeness that is tied to a target population.

8 Freedman (2006, p. 692) argues that such an urn model 'applies rather neatly to the as-if randomized natural experiments of the social and health sciences'.

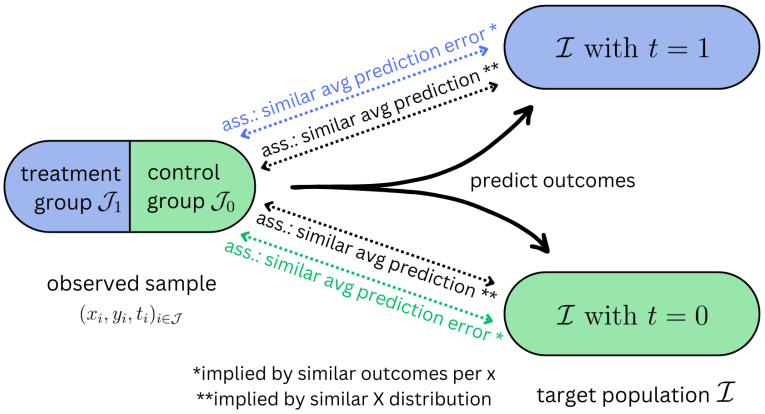

FIGURE 8.2: A schematic drawing of our framework: the observed sample is directly compared to the target sample, via an assumption regarding a similar distribution of covariates in $\mathcal{J}$ and $\mathcal{I}$ (ϵ-SAP) and an assumption that a predictor based on $\mathcal{J}_t$ is also calibrated on $\mathcal{I}$ under treatment t (δ-CTP).

As we will show now (and, more extensively, in Appendix 8.7.1), different estimators can be characterised by different predictors p, all with the goal of estimating the APO on the target sample though the average prediction on the observed/training sample. To make this work, two inductive assumptions are needed that tie the observed sample (and its treatment-wise groups) to the target sample (Figure 8.2). This dissolves the traditional separation into internal and external validity (which should be seen as an advantage, as we will argue in the next section).

For external validity, the standard framework typically assumes that the target data look like the observed data in the sense that they are sampled from the same marginal distribution of X_i (perhaps modulo a reweighing step); here, we instead require a more concrete and testable property: We assume that the average prediction on the observed population $\mathcal{J}$ is roughly the same as on the target population $\mathcal{I}$. We formalise this for $\epsilon > 0$ as the **ϵ-stable average predictions (ϵ-SAP)** assumption

$$\rho(\mathcal{I}, t) \approx_\epsilon \rho(\mathcal{J}, t). \tag{8.14}$$

where we denote the average prediction on observed and target sample by

$$\rho(\mathcal{J}, t) := \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} p(x_i, t) \qquad \text{and} \qquad \rho(\mathcal{I}, t) := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(x(i), t),$$

respectively. In comparison to the conventional assumption that covariates in training and target population are sampled from the same distribution (which

can be relaxed with reweighing methods), ϵ-SAP depends on the predictor p and is weaker (Remark 8.3 in Section 8.4.1). Another important difference for accountability is that whether it held in a given application can be directly read off the available data after deployment, without further assumptions.

In addition to ϵ-SAP, we need our predictor to work well on the target population. This assumption is formalised as **δ-calibration on the target population (δ-CTP)**,

$$\mu(\mathcal{I}, t) \approx_\delta \rho(\mathcal{I}, t). \tag{8.15}$$

Essentially, this says that the errors of our predictor, when applied to the target population, will roughly average out.[9] Together, these two assumptions allow us to make an $\epsilon + \delta$-good approximation of the APO under treatment t:

$$\mu(\mathcal{I}, t) \approx_\delta \rho(\mathcal{I}, t) \approx_\epsilon \rho(\mathcal{J}, t). \tag{8.16}$$

Different approaches to causal inference provide different estimators of the APO $\mu(\mathcal{I}, t)$ through different predictors p; in every case, the δ-CTP assumption needs to be justified individually. Typically, p is calibrated on $\mathcal{J}_t$ for all $t \in \mathcal{T}$ by design, so that δ-CTP follows if we assume that the calibration error of p on $\mathcal{I}$ is δ-close to the calibration error on $\mathcal{J}_t$:

$$\rho(\mathcal{I}, t) - \mu(\mathcal{I}, t) \approx_\delta \sum_{i \in \mathcal{J}_t} p(x_i, t_i) - \sum_{i \in \mathcal{J}_t} y_i = 0. \tag{8.17}$$

For example, we can see RCT-based inference as a using the degenerate predictor that predicts the group-wise mean for each individual, i.e.

$$p : (x, t) \mapsto \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} y_t. \tag{8.18}$$

Then

$$\rho(\mathcal{I}, t) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(x(i), t) = \frac{|\mathcal{I}|}{|\mathcal{I}|} \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} y_t, \tag{8.19}$$

s.t. δ-CTP then boils down to the assumption that the mean outcome in $\mathcal{J}_t$ is δ-close to the mean outcome in $\mathcal{I}$ under treatment t, which we have discussed above. A more interesting case is matching or inverse probability weighing.

9 Note that this uses calibration in the general sense of Dawid (2017) and Höltgen and Williamson (2023), that is, without conditioning on predictions.

8.3.4 *Observational setting and matching*

One of the most basic techniques is **exact matching** (Rosenbaum and Rubin, 1983). To capture this, define covariate-wise subgroups $\mathcal{I}^x, \mathcal{J}^x, \mathcal{J}_t^x$ for $x \in \mathcal{X}, t \in \mathcal{T}$,

$$\mathcal{I}^x := \{i \in \mathcal{I} \ : \ x(i) = x\}, \tag{8.20}$$

$$\mathcal{J}^x := \{i \in \mathcal{J} \ : \ x_i = x\}, \tag{8.21}$$

$$\mathcal{J}_t^x := \{i \in \mathcal{J}_t \ : \ x_i = x\}. \tag{8.22}$$

For a sufficiently coarse-grained set of covariates, we may observe all combinations (x, t) of covariates and treatments—which is often called 'positivity' or 'common support':

$$\forall x \in \mathcal{X}, t \in \mathcal{T} : \mathcal{J}_t^x \neq \emptyset. \tag{8.23}$$

Exact matching then uses the point-wise average predictor

$$p : (x, t) \mapsto \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i. \tag{8.24}$$

This predictor is much more fine-grained that the RCT predictor (8.18), as $p(x, t)$ predicts the observed x-*wise* average outcome in group $\mathcal{J}_t$, rather than averaging over all of $\mathcal{J}_t$. As we show in Proposition 8.2, this predictor satisfies δ-CTP (8.15) if the **average (signed) difference** between the x-wise mean outcomes

$$\mu_x(\mathcal{I}, t) := \frac{1}{|\mathcal{I}^x|} \sum_{i \in \mathcal{I}^x} y(i, t) \quad \text{and} \quad \mu_x(\mathcal{J}, t) := \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i,$$

is not strongly biased above or below zero, that is,

$$\left| \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \left(\mu_x(\mathcal{I}, t) - \mu_x(\mathcal{J}, t) \right) \right| < \delta. \tag{8.25}$$

Note that in the limit of infinite data drawn from some distribution, the common unconfoundedness assumption (8.5) would imply that the term $\mu_x(\mathcal{I}, t) - \mu_x(\mathcal{J}, t)$ goes to zero *for every* x with probability one; this is strictly stronger than (8.25), as the latter allows that the differences for different x values cancel each other out (see Section 8.4.1). We now show that in practice, (8.25) is indeed sufficient for the exact matching predictor; using the exact matching predictor for predicting the APO then amounts to inverse probability weighing (proof in Appendix 8.7.1):

Proposition 8.2 (Predicting the APO through matching)**.**
Fix any $t \in \mathcal{T}$. Assuming ϵ-SAP (8.14) and average (signed) difference below δ as in

(8.25), the exact matching predictor (8.24) gives us an $(\epsilon + \delta)$*-good approximation of the APO* $\mu(\mathcal{I}, t)$*:*

$$\left| \mu(\mathcal{I}, t) - \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{y_i}{e_t(x_i)} \right| < \epsilon + \delta, \tag{8.26}$$

where $e_t(x) := \frac{|\mathcal{J}_t^x|}{|\mathcal{J}^x|}$ *is the observed propensity score for treatment* t*.*

When the common support assumption (8.23) is not reasonable, one may be able to use a more coarse-grained approach. Using 'coarsened exact matching', we can predict averages not on every $x \in \mathcal{X}$ but on suitable subsets $U \in \Pi$ where Π is a partition of $\mathcal{X}$. We discuss this in Appendix 8.7.1, along with other methods such as doubly robust estimators, instrumental variables, and diff-in-diff.

8.4 COMPARING THE FRAMEWORKS

Every exposition of a new framework becomes clearer in comparison with established frameworks. Therefore, we now compare our proposition with standard RCMs more directly. We start with a descriptive comparison in mathematical form, relating the assumptions of ϵ-SAP and δ-CTP with i.i.d. sampling and unconfoundedness (Section 8.4.1). After that, we give a more subjective account of the practical advantages we see in the new framework (Section 8.4.2).

8.4.1 *Formal considerations*

We first note that ϵ-SAP follows from the conventional assumption that covariates in past and future are sampled from the same distribution, at least for bounded p and for large enough populations:

Remark 8.3 (average prediction in RCMs)**.**
For datapoints $x_1, ..., x_N$ *sampled i.i.d. from a distribution* P *that governs* X *on* $\mathcal{X}$ *and some bounded predictor* $p : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$*, it follows that* $\forall t \in \mathcal{T}$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} p(x_i, t) = \mathbb{E}_P[p(X, t)] \tag{8.27}$$

almost surely.

This means that if both datasets $\mathcal{J}$ and $\mathcal{I}$ are assumed to be sampled i.i.d. from the same distribution over $\mathcal{X}$, this implies that ϵ-SAP (8.14) holds almost surely in the limit of infinite data for any $\epsilon > 0$.

We can also relate the δ-CTP condition to the RCM framework; this can take different forms, as discussed in the context of different predictors, but often relies on the common unconfoundedness assumption,

$$Y_{ti} \perp\!\!\!\perp T_i \mid X_i. \tag{8.28}$$

It is often suggested that the unconfoundedness setting is 'probably the most important one in practice in the modern CI literature' (Imbens, 2020, p. 1163). Judea Pearl complains that 'I have yet to find a single person who can explain what [it] means in a language spoken by those who need to make this assumption or assess its plausibility in a given problem' (Pearl and Mackenzie, 2018, p. 281). Guido Imbens (2020) cites this passage and shoots back that 'simply assuming that one knows or can consistently estimate the joint distribution of all variables in the model' is also 'not helpful' (p. 1154). Imbens later adds that unconfoundedness 'is so common and well studied that merely referring to its label is probably sufficient for researchers to understand what is being assumed' (Imbens, 2020, p. 1164). To what extent it is well understood is not easy to settle, but it seems uncontroversial that 'the unconfoundedness assumption is not directly testable' (Imbens and Xu, 2024, p. 17). What our framework provides is an inductive assumption that can be tested in hindsight, which takes different forms for different predictors (see Section 8.3 and Appendix 8.7.1) and can be derived from unconfoundedness:

Remark 8.4 (conditional means under unconfoundedness)**.**
For a distribution P *over* $\mathcal{X} \times \mathcal{Y}_t \times \mathcal{T}$ *satisfying unconfoundedness (8.28) and datapoints* $(y_1, t_1), ..., (y_N, t_N)$ *sampled i.i.d. from the conditional distribution* $P(Y_{ti}, T_i|x)$*, we have, for* $\mathcal{J}_t^x(N) := \{1 \leqslant j \leqslant N : t_j = t\}$*,*

$$\lim_{N\to\infty} \frac{1}{|\mathcal{J}_t^x(N)|} \sum_{j \in \mathcal{J}_t^x(N)} y_j \;=\; \mathbb{E}_P[Y_i|X_i = x, T_i = t] \tag{8.29}$$

almost surely. This means that $\mu_x(\mathcal{J}, t)$ *converges to the conditional expectation for each* x *and* t*, which is also the limit to which* $\mu_x(\mathcal{I}, t)$ *converges (trivially).*

Thus, for $|\mathcal{J}_t^x|, |\mathcal{I}^x| \to \infty$, that the difference between $\mu_x(\mathcal{I}, t)$ and $\mu_x(\mathcal{J}, t)$ goes to zero such that (8.25) is easily satisfied (for any δ) with enough data. Note that unconfoundedness is indeed much stronger than (8.25) since the latter only requires that the signed average of the differences between the conditional means is small, whereas the remark implies that all *absolute* differences go to zero. Note that (8.25) also implies that δ-CTP is satisfied for the exact matching predictor, which justifies inverse probability weighing (Proposition 8.2).

8.4.2 *Practical considerations*

It is common practice to analyse to what extent the validity of results is sensitive to the violation of unconfoundedness in the form of unobserved confounders. While such sensitivity analyses can be useful, they still do not make the fundamental assumption of (approximate) unconfoundedness testable. Placebo tests go a bit further but rely on further untestable assumptions. A central aspect of the new framework is, then, that inductive assumptions can be formulated in ways that are directly testable and relatable to predictive success—while still allowing to use intuitions in terms of unconfoundedness to argue for the plausibility of the more concrete assumptions. Another assumption that is easily overlooked is the assumption of (i.i.d.) sampling which, as a relation between observables and a distribution, is difficult to even formalise. While the problem can be seen as less problematic for the pursuit of internal validity (as the distributions is per definition defined for the observed sample), this becomes more problematic for the question of generalisation or transferability to a target population. The new framework avoids this concept altogether (though it can be invoked as a limiting case, as in the related literature on non-probability sampling).

Perhaps the most fundamental novelty, however, is the shifted aim: from identification to prediction.[10] The notion of identification hinges on the notion of unobserved true parameters. Such parameters do not exist in our framework. This difference also has ramifications for the direct modelling of target populations, testability, the divide between internal and external validity. The lack of testability for assumptions like unconfoundedness has been discussed above. Our proposal is in the spirit of Freedman (1995, p. 33):

> There are statistical procedures for testing some of these assumptions. However, the tests often lack the power to detect substantial failures. Furthermore, model testing may become circular; breakdowns in assumptions are detected, and the model is redefined to accommodate. In short, hiding the problems can become a major goal of model building. Using models to make predictions of the future, or the results of interventions, would be a valuable corrective.

Our framework adds to this aim by specifying testable assumptions, which makes it possible to distinguish between potential sources of error and by making predictions the focus of modelling. This also entails that the target popula-

10 This view has predecessors even for the case of programme evaluation; e.g. Berk (1987, p. 184) notes that 'evaluations of program impact necessarily involve predictions' which 'involves expectations about the likely result under two or more conditions; they are predictions of the what-if variety.'

tion is explicitly included in the model, which requires (and allows) the modeller to directly grapple with the question of 'external validity'.[11] The crucial difference to work on transportability (Bareinboim and Pearl, 2013; Pearl and Bareinboim, 2014) is that the latter relies on strong modularity and stability assumptions that are not directly testable: it requires a full causal graph that is partially stable.

Another marked difference in our framework is that the separation between internal and external validity dissolves. This is a direct implication of abandoning the idea of identifying purported true parameters of abstract generative distributions. The concepts of external and internal validity are so engrained in social science research that this may seem extremely counter-intuitive to the experienced researcher. But also this idea is not completely novel: Indeed, the separate consideration of internal and external validity has recently been criticised as a reason for the neglect of external validity (Section 8.2.3) and the ensuing negative impact on public health research (Westreich et al., 2019). Furthermore, one can recover a version of internal validity by taking the sample population as the target population; we discuss this in the context of difference-in-difference estimators in Appendix 8.7.1.5.

Another, minor, difference in our framework is that now the APO, *vis-à-vis* the ATE, becomes the clear focus of the statistical enterprise. While some common estimators explicitly estimate the APOs as a first step, they are often understood as ATE estimators. In accordance with a view that many researchers already entertain, our framework more explicitly takes the APOs as the fundamental quantities, the ATE is not more than a formal comparison between APOs.

A more high-level difference is that the new framework explicitly works on averages rather than individuals. In RCMs, ATEs are typically seen as averages of individual treatment effects—indeed, the subscript i attached to all variables is meant to convey that the parameters and functions directly apply to each individual. Combined with high hopes for Machine Learning techniques (which use a similar formalism), there is an increasing push to 'fully personalized treatment effect estimates' (Athey and Imbens, 2016, p. 7353). This is despite the nature of statistics as a field capturing aggregate behaviour as well as cautionary voices pointing out the complexity of the social world (Section 8.2.2), where noise cannot be neatly controlled as in physical laboratories (Berk, 1987, p. 187).[12] In contrast, our framework dispenses not only with the notion individual effects but also discards the aim of estimating individual quantities; 'true' conditional

11 We are not claiming that it is impossible in principle to capture generalisation with conventional frameworks, just that it is less straightforward (Findley, Kikuta and Denly, 2021).

12 Sander Greenland notes that we 'cannot expect to discover numerically precise and general contextual laws analogous to those in physics' (Greenland, 2017, p. 4).

distributions are not defined and the goal is specified as predicting aggregate properties of outcomes in the target population.

8.5 TREATMENT RULES

So far, we have only considered APOs, that is, the *average* outcome if the *same* treatment is applied to everyone. This is enough to predict the average treatment effect, which is often considered the goal of causal inference. In many cases, however, we are not interested in applying the same treatment to everyone but instead want make treatment conditional on observed attributes. In this case, we may like to predict the average outcome of more complex covariate-based treatment rules $\pi : \mathcal{X} \to \mathcal{T}$,

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i, \pi(x(i))). \tag{8.30}$$

In this general formulation, assigning the same treatment rule to everyone, as considered so far, is captured by the degenerate policies $\pi_t : x \mapsto t$ for $t \in \mathcal{T}$. In recent years, starting with (Manski, 2004), there has been an increasing focus on learning more sophisticated treatment rules based on data. While we do not discuss the learning part in this paper, we note that our framework straightforwardly applies to predicting the average outcomes of such treatment rules.[13]

To apply our analysis to such treatment rules $\pi : \mathcal{X} \to \mathcal{T}$, it suffices to consider the sub-populations induced by the level sets of π, that is,

$$\mathcal{X}_t := \pi^{-1}(t) = \{x \in \mathcal{X} : \pi(x) = t\} \tag{8.31}$$

for $t \in \mathcal{T}$. Based on this, we partition $\mathcal{J}$ and $\mathcal{I}$ into subsets

$$\mathcal{J}^{\mathcal{X}_t} := \{i \in \mathcal{J} : x_i \in \mathcal{X}_t\} \quad \text{and} \quad \mathcal{I}^{\mathcal{X}_t} := \{i \in \mathcal{I} : x(i) \in \mathcal{X}_t\}, \tag{8.32}$$

then apply our machinery to all the pairs $\mathcal{J}^{\mathcal{X}_t}, \mathcal{I}^{\mathcal{X}_t}$ instead of $\mathcal{J}, \mathcal{I}$. For binary treatment $\mathcal{T} = \{0, 1\}$, this only means we consider two sub-populations instead of one population. The assumptions connecting $\mathcal{J}^{\mathcal{X}_t}$ and $\mathcal{I}^{\mathcal{X}_t}$ for each t are then analogous to those that we used in this paper to connect $\mathcal{J}$ and $\mathcal{I}$. For example, the assumption that two human populations $\mathcal{J}$ and $\mathcal{I}$ are similar in all relevant respects is hardly weaker than assuming that the respective sub-populations of (for example) people above 40 are similar, as are those below 40. However, this becomes stronger and stronger if we require this for more and more policies π

13 We prefer 'conditional' to 'individualised' to highlight the dependence of predictions on the choice of attributes taken into account.

simultaneously; requiring this for all possible policies would mean that the two populations must be exactly alike.[14]

Beyond deterministic treatment rules $\pi : \mathcal{X} \to \mathcal{T}$, one may also be interested in more general *stochastic treatment rules* $\pi : \mathcal{X} \to \Delta(\mathcal{T})$, where $\Delta(\mathcal{T})$ denotes the set of probability distributions over $\mathcal{T}$. For binary treatment decisions, that means that π assigns a treatment probability to each $x \in \mathcal{X}$. There are at least two distinct arguments for using such stochastic treatment rules: an ethical, and an epistemic one (Jain, Creel and Wilson, 2024). The ethical argument is that hard cut-offs are unfair because two people on opposite sides of the threshold are treated very differently (Vredenburgh, 2022). The epistemic argument is that we often do not have much data for some covariates, and if we then never assign some treatments to them, we will never gain that information. This can exacerbate inequality in the case of underrepresented groups, as discussed by O'Neil (2017) and analysed in a bandit setting by Li, Raymond and Bergman (2020). In causal inference, this resurfaces in terms of the assumptions we rely on when analysing data. With stochastic treatment rules, we get a well-defined propensity score by design that we can use for subsequent modelling. That is, we satisfy the assumption of 'missing at random' (Rubin, 1976) (see Footnote 24) *and* get direct access to the conditional probabilities. If the propensity score never reaches 0 or 100 per cent, we also satisfy positivity/common support. We also suggest calibration on sets of equal propensity as a testable criterion in Appendix 8.7.1. This suggests that it could be beneficial to use treatment rules which only assign a limited set of treatment probabilities, such as multiples of 10%, instead of the whole interval $[0, 1]$, in order to facilitate the analysis.

In general, for the assumptions to be testable in hindsight, we need to have access to the deployed treatment rule. This is difficult in settings where causal models are only used to assist decision making, e.g. as information offered to case workers. This is why it is particularly suitable to ADSs. As explained above, the chosen treatment rules do not have to be deterministic, as long as the 'weights' of the weighted lotteries are known. For simplicity and to directly compare with the bulk of the RCM literature, we have so far restricted ourselves to averages—but we do consider it important to go beyond this. In Appendix 8.7.2 we consider other properties of the potential outcome distribution.

14 Similarly, restricting the number of considered partitions is also necessary for risk minimisation (Vapnik, 1982), multi-calibration (Hébert-Johnson et al., 2018), and randomness (Derr and Williamson, 2024).

8.6 DISCUSSION

Causal inference methodology is increasingly becoming a hot topic in the literature on ADSs, mostly relying on Rubin's potential outcome framework (Liu et al., 2025; Zezulka and Genin, 2024) In this paper, we have presented an adaptation of this framework to model the affected populations directly, without relying on abstract distributions. The focus shifts from identifying theoretical properties of conjured distributions to predicting outcomes on the target population, while shedding new light on established methodologies rather than submitting them to the flames. In practice, developers can then directly predict the observable effects of policies and, by relying only on testable assumptions, retrospectively analyse sources of error in the modelling—which is also important for auditing. Assumptions and predictions that are otherwise formulated abstractly can now be put to scrutiny, such that developers can be more easily held accountable for design choices. We stress, however, that the formal framework, as any technical element, can never guarantee accountability by itself, as the latter critically depends on organisational processes and relationships (Cobbe, Lee and Singh, 2021; Metcalf et al., 2021).

While the focus on testability to enhance accountability appears novel, the project 'to reconfigure causal inference as the task of predicting what would happen under a hypothetical future intervention' (Dawid, 2022, p. 299) already has ardent advocates such as Philip Dawid (2000, 2021). Similar sentiments have also been expressed, for example Ragnar Frisch[15], Berk (1987), Greenland (2012)[16] or Hernán (2016, p. 679), who notes that

> The goal of the potential outcomes framework is not to identify causes–or to "prove causality", as it sometimes said. That causality cannot be proven was already forcibly argued by Hume in the 18th century. Rather, quantitative counterfactual inference helps us predict what would happen under different interventions.

These more interpretational considerations should, however, not divert attention from the practical benefits of explicitly modelling the target population and making testable assumption. Indeed, one may see the proposed framework either as a less metaphysical and abstract alternative to RCMs or simply as an

15 Already 'the founding father of modern econometric causal policy analysis' (Heckman and Pinto, 2024, p. 4) argued that 'the scientific [...] problem of causality is essentially a problem regarding our way of thinking, not a problem regarding the nature of the exterior world' (Frisch, 1930, p. 36).

16 Greenland here even believes to discern 'a subtle conceptual revolution that recognizes causal inference as a prediction problem' (p. 44).

empirically minded amendment. The framework is especially suitable for automated decision systems because i) it is often possible to know the treatment rule exactly (Section 8.5), ii) accountability is particularly relevant for consequential decisions about people, and iii) these systems are typically deployed for tasks where empirical data, but no reliable scientific knowledge is available which would warrant untestable claims about stable distributions and independence conditions (Section 8.2.2). By re-igniting discussions about the direct testability of all aspects of our models (Freedman, 1995), and by providing a constructive alternative to existing approaches, we hope to contribute to practices and systems that empower decision makers to develop trustworthy models, and fora and the public to hold decision makers accountable. While ADSs raise the stakes compared to human decision making through their scale, they *can* facilitate accountability as reviewability (Cobbe, Lee and Singh, 2021) through automatic documentation—but to make use of this opportunity, we need to establish adequate technical practices and accountability relationships.

8.7 APPENDIX

8.7.1 *New perspectives on established methodology*

Statisticians, economists, and others have developed various methods in the conventional RCM framework, many of which are well-established by now. In this appendix, we survey a few of them and demonstrate how we can recover them in our framework with weaker assumptions. The purpose of this is twofold: First, to showcase the new framework and demonstrate how it can capture and explain established methods, and second, to inspect these methods themselves and provide new perspectives that enhance our understanding of them.

We start with the proof of Proposition 8.2:

Proposition 8.2 (Predicting the APO through matching)**.**
Fix any $t \in \mathcal{T}$*. Assuming* ϵ*-SAP (8.14) and average (signed) difference below* δ *as in (8.25), the exact matching predictor (8.24) gives us an* $(\epsilon + \delta)$*-good approximation of the APO* $\mu(\mathcal{I}, t)$*:*

$$\left| \mu(\mathcal{I}, t) - \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{y_i}{e_t(x_i)} \right| < \epsilon + \delta, \tag{8.33}$$

where $e_t(x) := \frac{|\mathcal{J}_t^x|}{|\mathcal{J}^x|}$ *is the observed propensity score for treatment* t*.*

Proof. First, we get δ-CTP for predictor p from (8.25) via

$$\mu(\mathcal{I}, t) - \rho(\mathcal{I}, t) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i, t) - p(x(i), t) \tag{8.34}$$

$$= \frac{1}{|\mathcal{I}|} \sum_{x \in \mathcal{X}} \sum_{i \in \mathcal{I}^x} \left(y(i, t) - \sum_{j \in \mathcal{J}_t^x} \frac{y_j}{|\mathcal{J}_t^x|} \right) \tag{8.35}$$

$$= \frac{1}{|\mathcal{I}|} \sum_{x \in \mathcal{X}} |\mathcal{I}^x| \left(\sum_{i \in \mathcal{I}^x} \frac{y(i, t)}{|\mathcal{I}^x|} - \sum_{i \in \mathcal{J}_t^x} \frac{y_i}{|\mathcal{J}_t^x|} \right) \tag{8.36}$$

$$\approx_\delta 0. \tag{8.37}$$

Then

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i, t) \approx_\delta \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(x(i), t) \tag{8.38}$$

$$\approx_\epsilon \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} p(x_i, t) \tag{8.39}$$

$$= \frac{1}{|\mathcal{J}|} \sum_{x \in \mathcal{X}} |\mathcal{J}^x| \cdot p(x, t) \tag{8.40}$$

$$= \frac{1}{|\mathcal{J}|} \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i \tag{8.41}$$

$$= \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{y_i}{e_t(x_i)}, \tag{8.42}$$

where (8.39) uses ϵ-SAP (8.14). □

8.7.1.1 *Coarsened exact matching*

When the (strong) common support assumption is not reasonable, one may be able to use a more coarse-grained approach, that is, **coarsened exact matching**. On this approach, we predict averages not on every $x \in \mathcal{X}$ but on suitable subsets $U \in \Pi$ where Π is a partition of $\mathcal{X}$. For $U \in \Pi$, $t \in \mathcal{T}$, define

$$\mathcal{I}^U := \{i \in \mathcal{I} : x(i) \in U\} \tag{8.43}$$

$$\mathcal{J}_t^U := \{i \in \mathcal{J} : x_i \in U \wedge t_i = t\}. \tag{8.44}$$

Then we may use the predictor

$$p : (x, t) \mapsto \frac{1}{|\mathcal{J}_t^{U(x)}|} \sum_{i \in \mathcal{J}_t^{U(x)}} y_i, \tag{8.45}$$

(with $U(x)$ denoting the $U \in \Pi$ containing x) to predict the APO: Again, δ-CFD can be expressed as a signed average difference, now with the differences in subsets $\mathcal{I}^U$:

$$\left| \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^U|}{|\mathcal{I}|} \left(\mu_U(\mathcal{I}, t) - \mu_U(\mathcal{J}, t) \right) \right| < \delta. \tag{8.46}$$

Note that this is already satisfied if the differences between the average outcomes are δ-small for all U, i.e.

$$\forall U \in \Pi : \quad \frac{1}{|\mathcal{I}^U|} \sum_{i \in \mathcal{I}^U} y(i,s) \approx_\delta \frac{1}{|\mathcal{J}_t^U|} \sum_{i \in \mathcal{J}_t^U} y_i. \tag{8.47}$$

Also note that ϵ-SAP (8.14) for the coarsened matching predictor is satisfied if

$$\forall U \in \Pi : \left| \frac{|\mathcal{I}^U|}{|\mathcal{I}|} - \frac{|\mathcal{J}^U|}{|\mathcal{J}|} \right| < \frac{\epsilon \cdot |\mathcal{J}_t^U|}{\sum_{i \in \mathcal{J}_t^U} y_i}. \tag{8.48}$$

Proposition 8.5 (Predicting the APO through coarsened matching)**.**
Fix any $t \in \mathcal{T}$. Assuming ϵ-SP (8.14) for p as in (8.45) and δ-average signed difference as in (8.46), the coarsened exact matching predictor gives us an $(\epsilon+\delta)$-good approximation of the APO for t:

$$\left| \mu(\mathcal{I}, t) - \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{y_i}{e_{t,\Pi}(x_i)} \right| < \epsilon + \delta, \tag{8.49}$$

where $e_{t,\Pi}(U) := \frac{|\mathcal{J}_t^U|}{|\mathcal{J}^U|}$ and $e_{t,\Pi}(x) := e_{t,\Pi}(U(x))$, with $U(x)$ being the $U \in \Pi$ with $x \in U$.

Proof. First, we get δ-CFD from (8.46) via

$$\begin{aligned}
\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i,t) - p(x(i),t) &= \frac{1}{|\mathcal{I}|} \sum_{U \in \Pi} \sum_{i : x(i) \in U} \left(y(i,s) - \sum_{j \in \mathcal{J}_t^U} \frac{y_j}{|\mathcal{J}_t^U|} \right) && (8.50)\\
&= \frac{1}{|\mathcal{I}|} \sum_{U \in \Pi} |\mathcal{I}^U| \left(\sum_{i \in \mathcal{I}^U} \frac{y(i,s)}{|\mathcal{I}^U|} - \sum_{i \in \mathcal{J}_t^U} \frac{y_i}{|\mathcal{J}_t^U|} \right) && (8.51)\\
&\approx_\delta 0. && (8.52)
\end{aligned}$$

Then

$$\begin{aligned}
\mu(\mathcal{I}, t) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i,t) &\approx_\delta \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(x(i),t) && (8.53)\\
&\approx_\epsilon \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} p(x_i,t) && (8.54)\\
&= \frac{1}{|\mathcal{J}|} \sum_{U \in \Pi} \sum_{i \in \mathcal{J}^U} p(x_i,t) && (8.55)\\
&= \frac{1}{|\mathcal{J}|} \sum_{U \in \Pi} |\mathcal{J}^U| \cdot \frac{1}{|\mathcal{J}_t^U|} \sum_{i \in \mathcal{J}_t^U} y_i && (8.56)\\
&= \frac{1}{|\mathcal{J}|} \sum_{U \in \Pi} \frac{1}{e_{t,\Pi}(U)} \sum_{i \in \mathcal{J}_t^U} y_i && (8.57)\\
&= \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{y_i}{e_{t,\Pi}(x_i)}. && (8.58)
\end{aligned}$$

□

We can derive some insights from our analysis. The RCM literature typically assumes that there is a true underlying local propensity score that we can estimate, which allows for the construction of consistent estimators of the ATE. Outside of study designs that use weighted lotteries, it is not clear what the propensity score corresponds to in the real world; it is, thus, questionable if it makes sense to estimate this missing ground truth. If assumptions such as smoothness of the propensity score in $\mathcal{X}$ are made (which are implicit for any estimation method), it seems less problematic to explicitly assume 'constant propensity' on all $U \in \Pi$. This would already imply (8.47) (together with the common assumption that $\mathcal{I}$ and $\mathcal{J}$ are sampled from the same distribution). Note that our analysis is very similar to the propensity score theorem (Rosenbaum and Rubin, 1983) which derives $Y_{0i}, Y_{1i} \perp\!\!\!\perp T_i | e(X_i)$ from the unconfoundedness assumption $Y_{0i}, Y_{1i} \perp\!\!\!\perp T_i | X_i$: In our derivation, we weaken the unconfoundedness assumption to the statement that future t-outcomes are on average similar to the data we have for $\mathcal{J}_t$ and then show that we can 'condition on' sets of equal propensity. This can be seen as interpolating between exact matching and RCTs, where we basically assume that the propensity score is constant everywhere.

We have adopted the name 'coarsened exact matching' from (Iacus, King and Porro, 2012). They propose to coarsen variable-wise, e.g. applying a grid; we have shown that if we coarsen to sets that can be thought to have a constant propensity score, the predictions are well-founded. In line with this, it has also been suggested by Little (1986) to coarsen the considered covariates into groups of similar predicted propensity score to reduce variance, which they call 'response propensity stratification'. Kang and Schafer (2007) report that this indeed provides more robust estimates. As a last remark, one might say in line with the general theme of our approach, that predictions based on matching approaches do not match treatment to control group, but match observed groups to anticipated future populations.

8.7.1.2 *General calibrated predictors*

After having discussed specific matching predictors that can be expressed in terms of observed data, we now consider arbitrary $p : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$. An example for a wider class of predictors is given by the increasingly used Machine Learning algorithms.[17] Compared to matching, we now have less reason to believe in a low average signed error because learning algorithms often lead to systematic errors through their inductive biases. Analogous to coarsened matching, one

17 For now, we sidestep questions of overfitting by simply taking $\mathcal{J}$ to be a validation set.

may thus aim for low area-wise error on a partition Π of $\mathcal{X}$ where we can hope that the error will be similar on future data, in the sense of

$$\frac{1}{|\mathcal{I}^U|}\sum_{i\in\mathcal{I}^U} p(x(i),t) - y(i,t) \approx_\delta \frac{1}{|\mathcal{J}_t^U|}\sum_{i\in\mathcal{J}_t^U} p(x_i,t) - y_i. \tag{8.59}$$

Intuitively, this would be guaranteed (in the limit) on areas of 'constant propensity': In the distributional framework of RCMs, constant propensity on U would mean that the distribution of X on $\mathcal{J}_t^U$ is equal to that of $\mathcal{J}^U$ which is, in turn, equal to that on $\mathcal{I}^U$ – which means that, since $P(Y|X)$ is the same on $\mathcal{J}_t$ as on $\mathcal{I}$ via the unconfoundedness assumption, the joint distribution $P(X,Y)$ is the same on $\mathcal{J}_t^U$ as on $\mathcal{I}^U$. In settings where such distributions are difficult to justify, one may look for other ways to justify (8.59), which is, after all, much weaker than assumptions about propensity scores.

Proposition 8.6 (Predicting the APO through ML)**.**
Fix any $t \in \mathcal{T}$. Assuming ϵ-SP (8.14) for some p and (8.59) for all U in some partition of $\mathcal{X}$, we get an $(\epsilon+\delta)$-close approximation of the APO for t:

$$\left|\mu(\mathcal{I},t) - \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} p(x_i,t)\right| < \epsilon+\delta. \tag{8.60}$$

Proof.

$$\mu(\mathcal{I},t) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t) = \sum_{U\in\Pi}\frac{|\mathcal{I}^U|}{|\mathcal{I}|}\frac{1}{|\mathcal{I}^U|}\sum_{i\in\mathcal{I}^U} y(i,t) \tag{8.61}$$

$$\approx_\delta \sum_{U\in\Pi}\frac{|\mathcal{I}^U|}{|\mathcal{I}|}\frac{1}{|\mathcal{I}^U|}\sum_{i\in\mathcal{I}^U} p(x(i),t) \tag{8.62}$$

$$= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(x(i),t) \tag{8.63}$$

$$\approx_\epsilon \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} p(x_i,t). \tag{8.64}$$

□

8.7.1.3 *Doubly robust estimators*

As discussed above (see (8.25)), the calibration error on future data can be expressed in terms of the average x-wise (signed) prediction error:

$$\rho(\mathcal{I},t) - \mu(\mathcal{I},t) = \frac{1}{|\mathcal{I}|}\sum_{x\in\mathcal{X}}\sum_{i\in\mathcal{I}^x}(p(x,t)-y(i,t)) \tag{8.65}$$

$$= \sum_{x\in\mathcal{X}}\frac{|\mathcal{I}^x|}{|\mathcal{I}|}\left(p(x,t)-\mu_x(\mathcal{I},t)\right). \tag{8.66}$$

For the matching predictors, one may hope that this is close to zero through something akin to the unconfoundedness assumption, given that the predictor is simply the local observed average outcome (see Remark 8.4).

Alternatively, one may try to predict the local weights, using doubly robust estimators. These estimators use predictions p and w of both the mean and the propensity score; they correctly predict the APO when at least one of the two is correct. In our framework, the doubly robust estimator due to Robins and Rotnitzky (1995) works as follows.[18] The idea is to estimate the APO via

$$p_{dr}(x,t) := p(x,t) + w(x,t)\frac{1}{|\mathcal{J}^x|}\sum_{i\in\mathcal{J}_t^x}(y_i - p(x,t))\,. \tag{8.67}$$

Here, we need to assume either that

$$\sum_{x\in\mathcal{X}}\frac{|\mathcal{J}^x|}{|\mathcal{J}|}(\mu_x(\mathcal{I},t) - \mu_x(\mathcal{J},t))\cdot w(x,t) \approx_\delta 0, \tag{8.68}$$

or that

$$\sum_{x\in\mathcal{X}}\frac{|\mathcal{I}^x|}{|\mathcal{I}|}(\mu_x(\mathcal{I},t) - \mu_x(\mathcal{J},t)) \approx_\delta 0. \tag{8.69}$$

This is justified if the signed differences between x-wise averages can be assumed to be close to zero on average and/or not strongly correlated (as a function of x) with $w(x,t)$, similar to (8.25). This is, again, related to but weaker than the conventional unconfoundedness assumption.[19]

Then it is sufficient to either correctly estimate the conditional means, i.e.

$$\forall x\in\mathcal{X}, t\in\mathcal{T}: \quad p(x,t) = \mu_x(\mathcal{I},t), \tag{8.70}$$

or to correctly estimate how often each x occurs in J_t as compared to $\mathcal{I}$ through w in the sense of[20]

$$\forall x\in\mathcal{X}, t\in\mathcal{T}: \quad w(x,t) = \frac{|\mathcal{I}^x|}{|\mathcal{J}_t^x|}\frac{|\mathcal{J}|}{|\mathcal{I}|}. \tag{8.71}$$

Proposition 8.7 (Predicting the APO through doubly-robust estimators)**.** *Fix any $t\in\mathcal{T}$. Assuming ϵ-SP (8.14), the doubly robust estimator $p_{dr}(x,t)$ provides an $(\epsilon+\delta)$-good approximation of the APO for t in the sense of*

$$\left|\mu(\mathcal{I},t) - \sum_{x\in\mathcal{X}}\frac{|\mathcal{J}^x|}{|\mathcal{J}|}p_{dr}(x,t)\right| < \epsilon+\delta, \tag{8.72}$$

if either (8.68) and (8.70), or (8.69) and (8.71) hold.

18 Kang and Schafer (2007, p. 537) note that '[s]ome DR estimators have been known to survey statisticians since the late 1970s.'

19 Note that we typically assume that the marginal distribution over x is the same for train and test, such that the difference between (8.68) and (8.69) in that regard is not too important.

20 One can recover the RCM version of doubly robust estimators, simply using the inverse of the empirical propensity score $w(x,t) = \frac{|\mathcal{J}^x|}{|\mathcal{J}_t^x|}$, under the (strong) assumption that $\forall x\in\mathcal{X}: \frac{|\mathcal{J}^x|}{|\mathcal{J}|} = \frac{|\mathcal{I}^x|}{|\mathcal{I}|}$.

Proof. From (8.70) and (8.68) we get

$$\sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} w(x,t) \frac{1}{|\mathcal{J}^x|} \sum_{i\in\mathcal{J}_t^x} (y_i - p(x,t))$$

$$= \sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} w(x,t) \left(\frac{1}{|\mathcal{J}^x|} \sum_{i\in\mathcal{J}_t^x} y_i - \frac{1}{|\mathcal{I}^x|} \sum_{i\in\mathcal{I}^x} y(i,t) \right) \tag{8.73}$$

$$= \sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} w(x,t) \left(\mu_x(\mathcal{J},t) - \mu_x(\mathcal{I},t)\right) \tag{8.74}$$

$$\approx_\delta 0. \tag{8.75}$$

Therefore, using ϵ-SP (8.14) once again, we get

$$\sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} p_{dr}(x,t) = \sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} \left(p(x,t) + w(x,t) \frac{1}{|\mathcal{J}^x|} \sum_{i\in\mathcal{J}_t^x} (y_i - p(x,t)) \right) \tag{8.76}$$

$$\approx_\delta \sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} (p(x,t) + 0) \tag{8.77}$$

$$= \frac{1}{|\mathcal{J}|} \sum_{i\in\mathcal{J}} p(x_i,t) \tag{8.78}$$

$$\approx_\epsilon \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} p(x(i),t) \tag{8.79}$$

$$= \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} y(i,t) = \mu(\mathcal{I},t). \tag{8.80}$$

Alternatively, if (8.71) holds instead of (8.70), we can derive

$$\rho(\mathcal{J}, t) = \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} p_{dr}(x, t) \tag{8.81}$$

$$= \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} \left(p(x, t) + w(x, t) \frac{1}{|\mathcal{J}^x|} \sum_{i \in \mathcal{J}_t^x} (y_i - p(x, t)) \right) \tag{8.82}$$

$$= \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} \left(p(x, t) + \frac{|\mathcal{J}|}{|\mathcal{J}^x|} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} (y_i - p(x, t)) \right) \tag{8.83}$$

$$= \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} \left(p(x, t) - \frac{|\mathcal{J}|}{|\mathcal{J}^x|} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} p(x, t) + \frac{|\mathcal{J}|}{|\mathcal{J}^x|} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i \right) \tag{8.84}$$

$$= \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} p(x, t) - \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} p(x, t) + \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i \tag{8.85}$$

$$\approx_\epsilon \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i \tag{8.86}$$

$$= \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \mu_x(\mathcal{J}, t) \tag{8.87}$$

$$\approx_\delta \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \mu_x(\mathcal{I}, t) = \mu(\mathcal{I}, t). \tag{8.88}$$

The first approximation uses ϵ-SP, whereas the second approximation uses (8.68). □

In this sense, the estimator is doubly robust. But both (8.70) and (8.71) are clearly very strong assumptions. Double ML (Chernozhukov et al., 2018) assumes that we converge to this ideal eventually. But the goal of Machine Learning is to minimise average prediction loss, not to identify true conditional distributions. Hence, it seems that double-ML, as double robust estimators, is still limited in its applicability, and statements such as the following may apply to double-ML as well: 'There are papers suggesting that under some circumstances, estimating a shaky causal model and a shaky selection model should be doubly robust. Our results indicate that under other circumstances, the technique is doubly frail' (Freedman and Berk, 2008, p. 401).

8.7.1.4 *Non-compliance: Instrumental variables*

In many settings, no randomness or unconfoundedness assumption can be justified for the treatment assignment. Sometimes, an instrumental variable is available that can be seen as random and stands in a particular relationship to the treatment assignment of interest. A classic example is the draft lottery as an instrument for examining the effect of military service on earnings (Angrist,

1990). Consider then an instrumental variable with values in $\mathcal{Z} = \{0, 1\}$ and assume that we have a predictor $p : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$ satisfying δ-CTP (8.15) for z-wise (rather than t-wise) predictions, in the sense that

$$\forall z \in \mathcal{Z} : \quad \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i, z) \approx_\delta \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(z, x(i)). \tag{8.89}$$

This could be justified by one of the approaches discussed above. If we then have a setup where z is essentially random, as in a lottery, we can simply define a predictor $p : \mathcal{Z} \to \mathcal{Y}$ based on the average outcome per $\mathcal{J}_z$ as in RCTs, i.e.

$$\forall z \in \mathcal{Z} : \quad p(z) := \frac{1}{|\mathcal{J}_z|} \sum_{i \in \mathcal{J}_z} y_i \approx_\delta \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i, z). \tag{8.90}$$

Now further assume that

$$\forall z \in \mathcal{Z}, t \in \mathcal{T} : \quad \forall i \in \mathcal{I}_{tz} : y(i, z = z) = y(i, t = t), \tag{8.91}$$

where

$$\mathcal{I}_{tz} := \{i \in \mathcal{I} : t(i, z) = t\} \tag{8.92}$$

is unknown. This echoes the common assumption that z affects y only through t, called the 'exclusion restriction' (Angrist, Imbens and Rubin, 1996). In our framework, it means that, in considering predictions for y, our model treats interventions on t exactly as it does values of t when z is intervened upon (or assigned randomly in the setup). Note that our groups $\mathcal{I}_{tz}$ differ from the compliance groups in the LATE estimates (Imbens and Angrist, 1994) in that they a) partition only the future population $\mathcal{I}$ and b) do so in two pairs, sorted by which treatment $t \in \{0, 1\}$ they take, given assignment z: $\mathcal{I}_{00} \cup \mathcal{I}_{10} = \mathcal{I} = \mathcal{I}_{01} \cup \mathcal{I}_{11}$.

For a potentially novel result, we further assume what we shall call 'dominance', namely that

$$\sum_{i \in \mathcal{I}_{01}} y(i, t = 0) \leqslant \sum_{i \in \mathcal{I}_{01}} y(i, t = 1) \tag{8.93}$$

and

$$\sum_{i \in \mathcal{I}_{10}} y(i, t = 0) \leqslant \sum_{i \in \mathcal{I}_{10}} y(i, t = 1). \tag{8.94}$$

The intuition here is that for large enough groups, we are confident that the treatment has no negative effect on the average outcome. This is not always sensible and needs to be justified for each case – one needs to already have some qualitative understanding of the treatments. Based on this, we can derive the following result:

Proposition 8.8 (Lower bound on the ATE with IVs)**.**
Assume ϵ-SP (8.14) and δ-CFD (8.89) for some predictor p, as well as the exclusion restriction (8.91) and dominance (8.93), (8.94). Then we can lower-bound the ATE via

$$\mu(\mathcal{I}, t = 1) - \mu(\mathcal{I}, t = 0) \;\geqslant\; \rho(\mathcal{J}, z = 1) - \rho(\mathcal{J}, z = 0) - 2(\epsilon + \delta). \tag{8.95}$$

Proof. Using (8.93) and (8.89), we can derive

$$\begin{aligned}
\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(x(i), z = 1) &\approx_\delta \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, z = 1) && (8.96)\\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_{01}} y(i, t = 0) + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_{11}} y(i, t = 1) && (8.97)\\
&\leqslant \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_{01}} y(i, t = 1) + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_{11}} y(i, t = 1) && (8.98)\\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 1). && (8.99)
\end{aligned}$$

For (8.94), we analogously get

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(x(i), z = 0) + \delta \geqslant \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 0). \tag{8.100}$$

Hence, we can lower-bound the difference in APOs by

$$\begin{aligned}
&\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 1) - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 0) + 2\delta && (8.101)\\
&\quad\geqslant \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(x(i), 1) - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} p(x(i), 0). && (8.102)
\end{aligned}$$

With the usual ϵ-SP assumption, we thus get

$$\begin{aligned}
&\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 1) - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i, t = 0) + 2\epsilon + 2\delta && (8.103)\\
&\quad\geqslant \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} p(x_i, z = 1) - \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} p(x_i, z = 0). && (8.104)
\end{aligned}$$

□

In the special case where the instrument is assigned randomly and we can assume (8.90), we get the following.

Corollary 8.9 (Lower bound on the ATE with randomised IVs)**.**
Assume ϵ-SP (8.14) and (8.90), as well as the exclusion restriction (8.91) and dominance (8.93), (8.94). Then we can lower-bound the ATE via

$$\mu(\mathcal{I}, t = 1) - \mu(\mathcal{I}, t = 0) \;\geqslant\; \mu(\mathcal{J}, t = 1) - \mu(\mathcal{J}, t = 0) \;- 2(\epsilon + \delta). \tag{8.105}$$

Proof. We can use the above Proposition and simply insert p as in (8.90). Then we get

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t=1) - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t=0) + 2\epsilon + 2\delta \tag{8.106}$$

$$\geqslant p(1) - p(0) = \frac{1}{|\mathcal{J}_1|}\sum_{i\in\mathcal{J}_1} y_i - \frac{1}{|\mathcal{J}_0|}\sum_{i\in\mathcal{J}_0} y_i. \tag{8.107}$$

□

Two differences to the LATE methodology are worth noting: First, LATE estimates supposedly identify the average treatment effect on the group of 'compliers' (Angrist, Imbens and Rubin, 1996), which for us is the group $\mathcal{I}_{00} \cap \mathcal{I}_{11}$. For this to make sense, we would need to assume that these groups are well-defined even if the respective treatment is not assigned. This makes these statements metaphysically strong and unverifiable in principle. This is related to what Dawid (2000) calls 'fatalism', namely the assumption 'that the various potential responses Y_{ti}, when treatment t is applied to unit i, as predetermined attributes of unit i, waiting only to be uncovered by suitable experimentation' (p. 412, notation adapted). Considering the LATE estimates that are usually ascribed to compliers, he notes that 'it is only under the unrealistic assumption of fatalism that this group has any meaningful identity, and thus only in this case could such inferences even begin to have any useful content' (p. 413). We agree that these groups are not well-defined, as counterfactuals are not. Ascribing specific LATEs to supposedly fixed subgroups defined by counterfactuals thus relies on strong metaphysical assumptions; and typically, we are interested not in such subgroups (even if they were well-defined), but in the whole population (Deaton, 2009; Heckman and Urzua, 2010) – which is a practical reason to focus on lower bounds in the ATE of the whole population. For our approach, the outcomes also need not be well-defined before the treatment is assigned. The estimation of the ATE could not be directly tested, but it could be tested by randomly assigning two treatments to a future group and observing the population averages. Not even this is possible with LATE, as the group itself cannot be determined by observation.

A second diffference is that our assumptions (8.93) and (8.94) are very different to the typical 'monotonicity' assumption (Imbens and Angrist, 1994), according to which there is nobody with $t(i,z=1)=0$ but $t(i,z=0)=1$, i.e. for whom higher z leads to lower t. Even if LATEs would make sense, it would arguably be very strong to assume that *no* such person exists: it would not be enough that it increases the chance of treatment for everyone. In contrast, (8.93) and (8.94) say that higher t leads to higher y *on average*. Hence are not

only weaker but also less metaphysical, as they do not involve counterfactuals. Unfortunately, they are still untestable – which arguably makes IV approaches somewhat less reliable than other approaches – still we show how they can be used for estimating quantities on the population level.[21]

8.7.1.5 *Predicting the past: Difference-in-differences*

While the approaches so far have a clear forward-looking component, difference-in-differences, as well as synthetic control methods, are by design more backwards-looking. Ultimately, however, all such methods are meant to inform policy making. We already discussed in subsection 8.2.2 that the distribution framing can entice one to be overly optimistic about the generalisability of one's results. As pointed out by Deaton and Cartwright (2018) (in the context of RCTs), a focus on internal validity 'is sometimes incorrectly taken to imply that results of an internally valid trial will automatically, or often, apply "as is" elsewhere, or that this should be the default assumption failing arguments to the contrary' (p.10). In the following, we demonstrate how our framework makes sense of diff-in-diff approaches and, in doing so, directly involves the question of generalisation or induction (the case for synthetic control methods is similar).

We consider the following setting: Assume that there are three groups $\mathcal{G} = \{A, B, C\}$; we have data for group A and B and want to inform treatment decisions about group C. The data concerns two steps $\mathcal{S} = \{0, 1\}$; our data includes treatment $t = 1$ at step $s = 1$ for group A and treatment $t = 0$ at step $s = 1$ for group B. We will consider groups

$$\mathcal{J}^0_a,\ \mathcal{J}^1_{a1},\ \mathcal{J}^0_b,\ \mathcal{J}^1_{b0},\ \mathcal{J}^0_c,\ \mathcal{I}_c$$

(in the form $\mathcal{J}^s_{gt}$) with observed mean outcomes

$$\mu(\mathcal{J}^0_a),\ \mu(\mathcal{J}^1_a, t = 1),\ \mu(\mathcal{J}^0_b),\ \mu(\mathcal{J}^1_b, t = 0),\ \mu(\mathcal{J}^0_c).$$

At step $s = 0$, we do not need treatment indicators so we denote the people in the three groups by $\mathcal{J}^0_a, \mathcal{J}^0_b$ and $\mathcal{J}^0_c$. Lastly, $\mathcal{I}_c$ denotes the people of group C at step $s = 1$, for which we intend to make an informed treatment assignment.

We then assume that the group C is comparable to A and to B in terms of the difference of averages between time steps with the same treatment, i.e.

$$\mu(\mathcal{I}_c, t = 1) - \mu(\mathcal{J}^0_c) \ \approx_\epsilon\ \mu(\mathcal{J}^1_a, t = 1) - \mu(\mathcal{J}^0_a) \tag{8.108}$$

21 It has already been previously observed by Robins (1989) and Manski (1990) that – without assuming dominance – one can bound the APO from above and below if there is an upper and a lower bound on possible outcomes.

and

$$\mu(\mathcal{I}_c, t = 1) - \mu(\mathcal{J}_c^0) \approx_\delta \mu(\mathcal{J}_b^1, t = 0) - \mu(\mathcal{J}_b^0). \tag{8.109}$$

We can, thus, predict the APOs as

$$\mu(\mathcal{I}_c, t = 1) \approx_\epsilon \mu(\mathcal{J}_a^1, t = 1) - \mu(\mathcal{J}_a^0) + \mu(\mathcal{J}_c^0). \tag{8.110}$$

and

$$\mu(\mathcal{I}_c, t = 0) \approx_\delta \mu(\mathcal{J}_b^1, t = 0) - \mu(\mathcal{J}_b^0) + \mu(\mathcal{J}_c^0). \tag{8.111}$$

In comparison to standard accounts of diff-in-diff, we have directly included the group C that we want to make predictions about, rather than trying to infer supposed causal relationships in A or B. If we want to inform policymaking through our modelling, then we need to think about the target group C. There may, however, also be reasons to make (unverifiable) predictions about what would have happened in A under $t = 0$ or in B under $t = 1$. For this, we can use the model by inserting A or B for C. Note that we do not, strictly speaking, *learn* anything about what *would* have happened: We merely make a (potentially well-founded) prediction – a prediction that is unverifiable in principle. Still, it may provide an argument for or against other models: In the well-known case concerning the minimum wage in New Jersey and Pennsylvania (Card and Krueger, 1994), the interesting insight is that their model contradicts the general economics model that predicts lower employment for higher wages. While their analysis does not constitute *empirical* evidence against the general model, it can – if considered well-justified – still provide a strong reason to criticize it.

8.7.2 *Beyond averages*

In general, a social planner is interested in choosing the policy that maximises desirable properties of the distributions of outcomes. This gives a further reason to focus on treatment-wise potential outcomes rather than supposed treatment effects: Even if meaningful, the distribution of individual treatment effects would not allow us to infer other properties of the outcome distributions beyond the mean (Manski, 1996, p. 714). For simplicity and to directly compare with the bulk of the RCM literature, we have so far restricted ourselves to averages – but we do consider it important to go beyond this. In the following, we consider three alternatives to the APO, that is, other properties of the potential outcome distribution. This means we consider scalar predictions rather than probabilities for binary outcomes, since for binary outcomes, the average would specify the complete distribution. An example of such an outcome is income.

8.7.2.1 *Three alternatives*

A first quantity of interest is the proportion of individuals with an income above a certain threshold (such as the poverty line). This just collapses into the APO if we consider the binary outcome of whether an individual has an income above that threshold. Therefore, we do not discuss this property further.

A second potentially interesting property of the outcome distribution is the median or, more generally, quantiles. The target quantity is the q-quantile of the (empirical) target distribution for $q \in (0, 1)$, that is, the inverse $F^{-1}_{\mathcal{I},t}(q)$ of the cumulative distribution function

$$F_{\mathcal{I},t}(y) := \frac{|\{i \in \mathcal{I} : y(i,t) \leqslant y\}|}{|\mathcal{I}|}. \tag{8.112}$$

This is not necessarily well-defined so we make it more specific by defining

$$\xi^t_q := \min\{y \in \mathcal{Y} : F_{\mathcal{I},t}(y) \geqslant q\}. \tag{8.113}$$

In words, ξ^t_q is the lowest outcome threshold such that at least $q\%$ of the target population fall below it. It turns out that this problem can also be almost reduced to that of APOs—by fixing the value of the quantile ξ^t_q and then again considering the binary outcome of whether an individual has an income above that threshold. There are two caveats to this reduction. The first one is that we do not see the quantile ξ^t_q on the target population, so we need to consider the estimator of that quantile as our threshold. Still, by the above reasoning, we can argue that the proportion of people above the chosen threshold should remain similar. The second caveat is that small variation in the proportion may correspond to a large variation in the quantile if the differences between the outcomes around the threshold are high. This requires a further assumption, bounding this variation by requiring that for some function $\alpha : \mathbb{R} \to \mathbb{R}$, $\forall y \in \{y(i,t) : i \in \mathcal{I}\}$,

$$|F_{\mathcal{I},t}(y) - q| < \beta \quad \to \quad |y - \hat{\xi}^t_q| < \alpha(\beta). \tag{8.114}$$

While it is then analogous to the case of average outcomes (modulo the mentioned changes), these changes may not be as intuitive so we walk through the quantile case below.

A third and last type of quantity, that we want to at least hint at, is based on the notion of social welfare functions (SWFs). A class of SWFs take a set of outcomes and reduce them to a number, such as the average, but they may also pay attention to other aspects of the distribution. SWFs that is particularly

interesting are those that are rank-dependent and equality-minded (Kitagawa and Tetenov, 2021) and can be characterised as

$$W_\omega(\mathcal{I}, t) := \frac{1}{Z} \sum_{i=1}^{N} y_i \cdot \omega(F_{\mathcal{I},t}(y_i)), \tag{8.115}$$

where $Z := \sum_{i=1}^{N} \omega(F_{\mathcal{I},t}(y_i))$ is a normalising constant and ω is non-negative and monotonically increasing, i.e. assigning lower weights to higher outcomes based on their rank/percentile. The average corresponds to a constant ω; other SWFs are the minimum, that is, how the worst off are doing, and measures in between. The more complex characterisation of such outcomes suggests that it is in general more difficult to characterise and satisfy the assumptions required for predicting this with precision. While we consider this an important topic, we leave it for future work.

8.7.2.2 *More details on quantiles*

We now return to the prediction of quantiles of the outcome distribution,

$$\xi_q^t := \min\{y \in \mathcal{Y} : F_{\mathcal{I},t}(y) \geqslant q\}. \tag{8.116}$$

For RCTs where it is justified to assume no bias between sample and target population, we can simply take the observed quantile in group $\mathcal{J}_t$ as an estimator. This can be justified when $\mathcal{J}_t$ and $\mathcal{I}$ can be considered as making up the same population, which means that the proportion p of values below a certain threshold should be similar in both groups. We denoting the q-quantile in $\mathcal{J}_t$ as

$$\xi_q^{\mathcal{J}_t} := \min\{y \in \mathcal{Y} : F_{\mathcal{J}_t}(y) \geqslant q\}. \tag{8.117}$$

We denote

$$\gamma := |F_{\mathcal{J}_t}(y) - q| \tag{8.118}$$

which is measurable (and indeed we can tweak q for γ to be zero). Then we formulate the dummy outcomes

$$\tilde{y}_i := \mathbb{1}[y_i \leqslant \xi_q^{\mathcal{J}_t}] \quad \text{and} \quad \tilde{y}(i,t) := \mathbb{1}[y(i,t) \leqslant \xi_q^{\mathcal{J}_t}].$$

Then from the unbiasedness assumption

$$F_{\mathcal{I},t}(\xi_q^{\mathcal{J}_t}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{y}(i,t) \approx_\delta \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} \tilde{y}_i = F_{\mathcal{J}_t}(\xi_q^{\mathcal{J}_t}), \tag{8.119}$$

analogous to (8.12), we get that

$$F_{\mathcal{I},t}(\xi_q^{\mathcal{J}_t}) \approx_\delta F_{\mathcal{J}_t}(\xi_q^{\mathcal{J}_t}) \approx_\gamma q. \tag{8.120}$$

Then the bounded variability assumption (8.114) lets us bound

$$|\xi_q^{\mathcal{J}_t} - \xi_q^t| < \alpha(\delta + \gamma). \tag{8.121}$$

For non-RCT samples, we can use an approach similar to matching or inverse probability weighing. It has been previously observed that one can use a version of inverse probability weighing to estimate the cumulative outcome distribution. For both fine-and coarse grained (i.e. stratified) approaches, it is common to use parametric propensity score models, as in Zhang et al. (2012). In line with out previous discussion of exact matching (and our discussion of coarse-grained exact matching in the Appendix), we discuss the use of empirical propensity scores here. The idea is again to weight instances for treatment t (that is, in group $\mathcal{J}_t$) based on their occurrence of their covariates in the entire sample $\mathcal{J}$. The estimator $\hat{\xi}_q^t$ is then defined as

$$\hat{\xi}_q^t := \min\left\{ y \in \mathcal{Y} : \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{\mathbb{1}[y_i \geqslant y]}{e_t(x_i)} \geqslant q \right\}, \tag{8.122}$$

where $e_t(x) := \frac{|\mathcal{J}_t^x|}{|\mathcal{J}^x|}$ is again the observed propensity score for treatment t. Now we define dummy outcomes w.r.t. $\hat{\xi}_q^t$, as

$$\tilde{y}_i := \mathbb{1}[y_i \leqslant \hat{\xi}_q^t] \quad \text{and} \quad \tilde{y}(i,t) := \mathbb{1}[y(i,t) \leqslant \hat{\xi}_q^t].$$

Similar to above, we define

$$\gamma := \left| \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}_t} \frac{\tilde{y}_i}{e_t(x_i)} - q \right|, \tag{8.123}$$

which is usually small. Then, similar to Section 8.3.4, we assume that the **average (signed) difference** between the x-wise proportions of outcomes below $\hat{\xi}_q^t$,

$$r_x^q(\mathcal{I}, t) := \frac{1}{|\mathcal{I}^x|} \sum_{i \in \mathcal{I}^x} \tilde{y}(i,t) \quad \text{and} \quad r_x^q(\mathcal{J}, t) := \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} \tilde{y}_i,$$

is not strongly biased above or below zero, that is,

$$\left| \sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} \left(r_x^q(\mathcal{I}, t) - r_x^q(\mathcal{J}, t) \right) \right| < \delta. \tag{8.124}$$

Analogous to ϵ-SAP, we assume that the distribution on $\mathcal{X}$ is similar on $\mathcal{J}$ compared to $\mathcal{I}$ in the sense that

$$\sum_{x \in \mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} r_x^q(\mathcal{J}, t) \approx_\epsilon \sum_{x \in \mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} r_x^q(\mathcal{J}, t). \tag{8.125}$$

Now since

$$F_{\mathcal{I},t}(\hat{\xi}^t_q) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tilde{y}(i,t) = \sum_{x\in\mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} r^q_x(\mathcal{I},t), \tag{8.126}$$

and

$$\sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} r^q_x(\mathcal{J},t) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}_t} \frac{\tilde{y}_i}{e_t(x_i)} \approx_\gamma p, \tag{8.127}$$

by definition of γ, this gives us

$$F_{\mathcal{I},t}(\hat{\xi}^t_q) \approx_\delta \sum_{x\in\mathcal{X}} \frac{|\mathcal{I}^x|}{|\mathcal{I}|} r^q_x(\mathcal{J},t) \approx_\epsilon \sum_{x\in\mathcal{X}} \frac{|\mathcal{J}^x|}{|\mathcal{J}|} r^q_x(\mathcal{J},t) \approx_\gamma p. \tag{8.128}$$

As above, via (8.114) we can then bound

$$|\hat{\xi}^t_q - \xi^t_q| < \alpha(\epsilon+\delta+\gamma). \tag{8.129}$$

8.7.3 *Linear regression*

So far, we have focussed on causal inference for binary treatments, as particularly relevant for programme evaluation, but it can also be applied to settings where treatment can take multiple values. The most popular model for such settings is linear regression. In their landmark textbook, Angrist and Pischke (2009, p. 52) vaguely note that '[a] regression is causal when the [true distributional model] it approximates is causal'. Regression for causal inference is slightly more controversial than for binary treatments, as it needs stronger assumptions that are nevertheless less visible. For example, Michael Freedman[22] notes that

> 'Lurking behind the typical regression model will be found a host of such assumptions; without them, legitimate inferences cannot be drawn from the model. There are statistical procedures for testing some of these assumptions. However, the tests often lack the power to detect substantial failures.' (Freedman, 1995, p. 33)

These caveats become more evident when showing how linear regression fits into our framework.

8.7.3.1 *Identifying correct models*

We start by demonstrating that, assuming there is a correct linear model, regression can identify this model. In our framework, this means that we assume there is a linear model

$$p(x,s) = a^*x + \beta^*s + c^* \tag{8.130}$$

22 See also his critique of causal regression in (Freedman, 2006, 2008).

that can describe the future average well for every $x \in \mathcal{X}$ (analogous to assuming a linear CEF) in the sense that

$$\forall x \in \mathcal{X}, t \in \mathcal{T}: \quad \frac{1}{|\mathcal{I}^x|} \sum_{i \in \mathcal{I}^x} y(i,s) \approx_\epsilon \frac{1}{|\mathcal{I}^x|} \sum_{i \in \mathcal{I}^x} a^* x + \beta^* s + c^*. \tag{8.131}$$

One sufficient assumption is that the $\mathcal{J}_t^x$ are comparable with the $\mathcal{I}^x$ in the sense that the residuals are not biased, i.e.

$$\forall x \in \mathcal{X}, t \in \mathcal{T}: \quad \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i - p(x,t) \approx_\delta \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}^x} y(i,t) - p(x(i),t) \tag{8.132}$$

Then clearly

$$\forall x \in \mathcal{X}, t \in \mathcal{T}: \quad \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} y_i \approx_{\epsilon+\delta} \frac{1}{|\mathcal{J}_t^x|} \sum_{i \in \mathcal{J}_t^x} a^* x + \beta^* t + c^*. \tag{8.133}$$

Alternatively, if the x_i are uncorrelated with the t_i in our data, then to show (8.133), it is also sufficient – instead of (8.132) – to assume only t-wise comparability, i.e.

$$\forall t \in \mathcal{T}: \quad \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i,t) - \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p(x(i),t) \approx_\delta \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} y_i - \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} p(x_i,t). \tag{8.134}$$

In this case, note that (8.131) implies

$$\begin{aligned} \forall s,t \in \mathcal{T}: & \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i,s) - \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} y(i,t) && (8.135)\\ & \approx_{2\epsilon} a^* x(i) + \beta^* s + c^* - a^* x(i) + \beta^* t + c^* && (8.136)\\ & = \beta^* \cdot (s-t) && (8.137) \end{aligned}$$

and thus

$$\forall s,t \in \mathcal{T}: \frac{1}{|\mathcal{J}_s|} \sum_{i \in \mathcal{J}_s} y_i - \frac{1}{|\mathcal{J}_t|} \sum_{i \in \mathcal{J}_t} y_i \approx_{2\delta+2\epsilon} \beta^* \cdot (s-t), \tag{8.138}$$

In the case of perfect fits, i.e. $\epsilon = \delta = 0$ fitting a linear model with MSE loss would identify the true model, as MSE elicits the mean. Furthermore, in the case of the x_i being uncorrelated with the t_i, the OVB formula tells us that the β estimator in the long regression is equal to the estimator in the regression $y = \beta \cdot t$, such that either of them would identify the true parameter. The fact that the assumptions in this subsection are very strong can be connected back to the cited critique by David Freedman – despite the identification and validity results for linear regression that can be derived in expectation, or in the limit of infinite data.[23]

23 It is also worth noting that linear models are often chosen less because there are good reasons to believe in linear models and more because they are nice to work with.

8.7.3.2 *Instrumental variables*

Here, we assume the assignment of the instrument z was 'random' in the sense that

$$\forall z \in \mathcal{Z}: \quad \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,z) \approx_\delta \frac{1}{|\mathcal{J}_z|}\sum_{i\in\mathcal{J}_z} y_i. \tag{8.139}$$

Now further assume (as usual for IV models) that z affects y only through t ('exclusion restriction') in the sense that

$$\forall i \in \mathcal{I}, z \in \mathcal{Z}: \quad y(i, z = z) = y(i, t(i,z)). \tag{8.140}$$

Assume further that the average outcome y of an intervention on t depends on t only via its average, i.e. that for any treatment assignment $\tau, \tau' : \mathcal{I} \to \mathcal{T}$ if it holds that

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tau(i) \approx_\gamma \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tau'(i) \tag{8.141}$$

implies

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,\tau(i)) \approx_\epsilon \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,\tau'(i)). \tag{8.142}$$

This means in particular that outcomes are on average linear in the treatment – which is also part of the conventional assumption that there is a true causal linear model $y = a^* x + \beta^* t + c$.

Then for any treatment rule $\tau_z : \mathcal{I} \to \mathcal{T}$ that roughly leads to the same average treatment as assigning z would do, i.e.

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tau_z(i) \approx_\gamma \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} t(i,z), \tag{8.143}$$

we know via (8.142), (8.140), and (8.139) that we can predict the average outcome of a treatment rule τ_z as

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,\tau_z(i)) \approx_\epsilon \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,t(i,z)) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y(i,z) \approx_\delta \frac{1}{|\mathcal{J}_z|}\sum_{i\in\mathcal{J}_z} y_i. \tag{8.144}$$

To use this, we need to know which instrument would have led to a similar average treatment. For this, we can investigate the data under the assumption that

$$\forall z \in \mathcal{Z}: \quad \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} t(i,z) \approx \frac{1}{|\mathcal{J}_z|}\sum_{i\in\mathcal{J}_z} t_i, \tag{8.145}$$

justified by the random assignment of z.

This leads to a procedure that is somewhat reminiscent of the two steps in 2SLS: To estimate the APO of a treatment t, we first analyse the observed relationship between instrument and treatment to find a $z \in \mathcal{Z}$ with (8.143). Then we use the observed relationship between instrument and outcome to predict the APO of the treatment. The main difference to 2SLS is that we do not try to identify parameters of an assumed linear model here, which makes it more general. Also note that our approach can be straightforwardly generalised to predict the APO of a policy $\iota : \mathcal{X} \to \mathcal{T}$ instead of a constant treatment t in cases where we have access to covariates $\mathcal{X}$.

8.7.4 *Connection to non-probability sampling*

Considering a treatment group $\mathcal{J}_t$ as a subsample of the observed sample $\mathcal{J}$ connects it to problems of survey sampling or missing data. Causal inference has been considered a missing data problem from the start by Rubin, see e.g. (Little and Rubin, 2020; Rubin, 1976). The connections are particularly clear in finite population settings. Inverse probability weighing was developed for finite population survey sampling (Horvitz and Thompson, 1952) – for cases when the propensity score is known by design.[24] Abadie et al. (2020) compute standard errors for estimating causal population means in terms of the uncertainty introduced by the random assignment of treatment in the observed data. They assume that the counterfactual outcomes are well-defined, but it should be possible to apply similar reasoning to the account pursued here. While survey sampling is not mentioned in the cited work, the authors do draw that connection in an earlier version (Abadie et al., 2014). Conversely, the non-probability survey sampling literature in particular (Elliott and Valliant, 2017; Wu, 2022) draws heavily on early work by Rubin and others. This literature is concerned with inference from non-representative samples from finite populations – more precisely, from 'samples without an identified design probability construct' (Meng, 2022, p. 341). Nowadays these problems are mostly discussed independently, with some exceptions. Mercer et al. (2017) explicitly construes non-probability survey sampling as a causal inference problem.

We go the opposite route and suggest seeing causal inference as an instance of predictions under non-probability sampling. Kang and Schafer (2007) note that 'the methods described in this article [for estimating a population mean from incomplete data] can be used to estimate an average causal effect by ap-

24 The strength of the assumption that there is a well-defined propensity score – part of the 'missing at random' assumption in (Rubin, 1976) – seems to be hardly discussed anymore, as it directly follows from the way the problem is set up.

plying the method separately to each potential outcome' (p. 525). In contrast to this and the above-mentioned work by Abadie and colleagues, however, we see the aim in making treatment-wise predictions rather than inferences about counterfactual outcomes or individual effects. That means that there is no so-called 'fundamental problem of causal inference' (Holland, 1986): knowing two mutually exclusive potential outcomes for some input would not help to make predictions. The basic idea is to build predictors which take treatment as just another attribute, but one whose distribution may change dramatically in the future. This means that we can see causal inference methods as treatment-wise predictors of potential outcomes (Figure 8.2).

ACKNOWLEDGEMENTS

We would like to thank Sebastian Zezulka, Michael Knaus, and Henri Pfleiderer for valuable feedback on earlier drafts.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645 as well as the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Benedikt Höltgen.

9

A THOUGHTFUL NARRATIVE EMERGES

Once men turned their thinking over to machines
in the hope that this would set them free. But that only
permitted other men with machines to enslave them.

— Frank Herbert, *Dune*

AUTHOR CONTRIBUTIONS

This is a single-author paper.

9.1 INTRODUCTION

The spectre that is haunting both scientific and public imagination—the spectre of Artificial Intelligence (AI)—exhibits a subtle tension within its most common narratives. On the one hand, AI is envisioned as a (potentially conscious) person. This applies not only to appearances in fiction but also to scientific debates about agency and the ascription of 'beliefs', 'intentions', and 'goals'. Such anthropomorphisms are applied not only to potential future systems but already to current ones, and suggest that such a system has a point of view. On the other hand, AI is envisioned as overcoming human subjectivity, idiosyncrasies, and bounded perspectives. Unconstrained by the limits of human physiology and psychology, AI strives towards flawlessness and objectivity. Taken together, AI promises to be a non-human human, a person without personality;[1] or, as Sabina Leonelli (2025) puts it, 'the answer to (some) humans' wish to transcend their [perceived] limits'.

In her admirable article, Leonelli develops an alternative narrative for intelligence and the future of technology under the name of Environmental Intelligence (EI). Her vision for digital technology drops both ideas of personhood and objectivity. The vision is not to create a universal uber-human but to build

1 Meta (2025) recently coined the term 'personal superintelligence', similar to a personal assistant.

technology in the service of human and non-human life on earth. It is instructive to compare this narrative with two other prominent and recently proposed visions for rebranding AI as a technology in the service of humankind.

First, the vision advocated by Machine Learning (ML) pioneer Yoshua Bengio is a 'Scientist AI', 'designed to explain the world from observations, as opposed to taking actions' (Bengio et al., 2025, p. 1). The main difference to more common narratives of AI is, thus, that the AI is not able to perform actions in the world by itself. The idea is that without real agency, AI systems can be more easily controlled by humans and put to good use, by answering questions and generating hypotheses—vaguely reminiscent of Douglas Adams' fictional supercomputer Deep Thought (except faster). While the label 'scientist' still suggests a person-like entity, it is explicitly designed not to have intentions and goals, making it less anthropomorphic. Second, Acemoglu and Johnson (2023) go a step further in 'Power and Progress' by suggesting to shun the framing of 'intelligence' altogether. They propose to shift the narrative from 'machine intelligence' to 'machine usefulness', which rejects anthropomorphisms and focuses on how technology can be beneficial for humans on individual tasks. In their view, the focus should be on how technology can 'complement human capabilities and empower people' (p. 299), rather than imitate or replace them.

Leonelli (2025) agrees with both these visions that the goal should be to create technology at the service of humans rather than independent entities. In contrast to them, she neither keeps the notion of intelligence as it is, nor does she shun it—instead, she defines it as an emergent capacity to interact adaptively with ones environment—a notion which, in her view, cannot be 'fully emulated or substituted by technology'. This perspective allows her to build a narrative that directly competes with common ideas of AI, collecting both normative and epistemic criticisms. Her notion of EI is different to the other two visions in that it explicitly emphasises the need for considering non-human life and climate change, based on both ethical and pragmatic considerations. But another interesting difference, which I shall focus on now, is her inclusion of transdisciplinary epistemic considerations. These are more integrated into the technological narrative, compared to Acemoglu and Johnson's focus on economics, and more cautious than Bengio et al.'s focus on algorithms (although the former also point to social and situational aspects of intelligence, p. 314). More specifically, what Leonelli's narrative does particularly well is expose the sterility of standard visions of AI.

9.2 STERILE INTELLIGENCE

Machine Learning systems are typically developed in rather sterile settings, lacking a direct interaction with the world. This includes many of the most impressive recent breakthroughs, such as games: Chess, Go, and Atari worlds all involve a discrete and comparatively simple setup. This falls under what has been dubbed 'Moravec's paradox', the observation that many tasks that are simple for humans are difficult for computers and vice versa.[2] Computers or ML systems are good at tasks in sterile environments because they can throw a lot of compute at well-structured problems with an unambiguous feedback signal. This makes it possible to train a Go champion in a matter of hours—which is still impressive. Another common example are curated datasets for all varieties of prediction tasks from image recognition to protein structures.

This sterility issue also applies to the most prominent class of models over the last years, large language models (LLMs). A particular focus in the last years has been given to mathematical reasoning, resulting in regular news about new milestones in mathematical capabilities (typically accompanied by a description of the kind of human that the models are currently on par with). Taking Leonelli's perspective, it is, however, quite clear that this provides no straightforward path to Environmental Intelligence. As put by John von Neumann: 'if people do not believe that mathematics is simple, it is only because they do not realize how complicated life is' (Alt, 1972). Within formal systems, there are simply significantly fewer degrees of freedom, which is something that sterile visions of AI typically ignore. It applies to LLMs more generally, as these models live in a world of tokens, a world of shadows as in Plato's allegory of the cave.[3] It is still much less complex than the natural or the social world—which is why the (fictional) supercomputer Deep Thought recommended building planet Earth for solving particularly difficult questions.

There is another aspect that makes mathematics 'easier' for ML systems: Recent advances in reasoning capabilities, such as in DeepSeek-R1-Zero, are based on the application of pure reinforcement learning to 'tasks such as coding, mathematics, science, and logic reasoning, which involve well-defined problems with clear solutions' (Guo et al., 2025, p. 10). This is similar to self-training in games

2 Moravec (1988, p. 15) noted already in the 80s that 'it has become clear that it is comparatively easy to make computers exhibit adult level performance in solving problems on intelligence tests'.

3 As put in a similar context by Birhane and McGann (2024, p. 9), 'You must go to play in the LLM's virtual world; *it cannot come to yours*' (their emphasis). This is not to say that current AI models navigating digital worlds cannot already do a lot of collateral damage in the real world, such as token generators let loose on social media or talking to humans.

that leads to shorter training times as mentioned above, and different from other LLM tasks which require reinforcement learning from human feedback (RLHF); RLHF was developed for tasks that 'do not have unambiguous ground truth' (Ziegler et al., 2020). Such tasks are much more cumbersome not only for training, but also for evaluation: What counts as a viable solution depends on human judgment. This lack of a solid ground truth is an issue that is increasingly surfacing in ML research, yet still not widely appreciated; Leonelli's framing of Environmental Intelligence brings it more to the forefront.

9.3 SHAKY GROUND TRUTHS

The observed lack of 'unambiguous ground truth' is considered '[u]nlike many tasks in ML' (Ziegler et al., 2020). However, this problem is much more widespread than commonly acknowledged. Arguably, most probabilistic prediction tasks do not have a clear ground truth. In sequence prediction with LLMs, for example, the conditional probability for the next token is not well-defined for every sequence: if a given sequence has never been uttered before, or only by a person at a very different place and time, how do we define that probability? The same holds for many other ML tasks. In contrast, the AI narrative of ever more data and deeper learning promises to get closer and closer to an imagined ground truth; for example, Bengio et al. (2025) envision that the Scientist AI will be 'trained by optimizing a training objective over possible explanations of the observed data which has a single optimal solution to this computational problem' and '[t]he more computing power ("compute") is available, the more likely it is that this unique solution will be approached closely'. Going beyond Leonelli's discussions of data, it is instructive to consider how data is typically conceptualised in ML, namely, as proxies for an underlying distribution. As exemplified by the main Deep Learning textbook, data is increasingly seen as a sort of mediator in the quest 'to match the true data-generating distribution p_{data}' (Goodfellow, Bengio and Courville, 2016, p. 130). Such limited perspectives on the workings of AI systems ignore two problems: First, that there is a choice which distribution to look at and second, that these distributions are a construct which can be more or less well-defined.

The first problem is basically the problem Leonelli (2025) discusses under premise 2: human judgment is used for determining what information we take into account for decision making. In terms of the 'data-generating distribution' over input features and target labels, this means the choice of input features and the choice of target labels. The former is the choice of abstraction or idealisation, which is needed to condense real-world events into computable chunks of in-

formation, whereas the latter is the choice of a decision-relevant target. While there is work on both choices in Machine Learning and Statistics, most research takes both for granted and assumes input and label space as given in the problem formulation.[4] Leonelli rightly emphasises the dependence on 'assumptions around what counts as the relevant background knowledge, analytic skills, and goals to be fulfilled through the data analysis at hand'. Indeed, the choice of 'ontology' (i.e. categories, quantities to measure, ...) in science depends on the specific goals (Danks, 2015) and non-epistemic values (Ludwig, 2016), and is a central part of science. In this sense, the Scientist AI vision, taking data for granted, arguably has a too narrow idea of science and a sterile notion of intelligence. Leonelli herself is one of the foremost experts on this topic, given her admirable range of work on 'data journeys'.

The second problem is that even taking an input space $\mathcal{X}$ and a target label space $\mathcal{Y}$ as fixed, the purported 'true data-generating distribution p_{data}' is not just out there in the world for us to find. In ML and Statistics, we are trained to see stable data-generating processes everywhere, but there are many settings where this is too much of an idealisation. This is particularly clear in social prediction tasks, such as predicting credit default or unemployment (Höltgen and Williamson, 2026b): Especially if there are few datapoints per $x \in \mathcal{X}$, then the 'true' conditional probability $P(Y|X)$ depends on the considered time frame: if there are only a few people with the characteristics x, then the empirical $P(Y|X = x)$ can strongly depend, for example, on whether one looks at the last ten or twenty years. This is because these ratios are not stable over time for many prediction tasks (Ding et al., 2021; Munger, 2023; Vela et al., 2022). The 'ground truth' probabilities also depend on the population to be considered: Is the conditional probability defined by similar people living in the same neighbourhood, city, or country? These are fundamental problems that also apply to other types of tasks and also relate to Leonelli's second premise. Collecting more data in such unstable settings often means that the target is moving—contrary to the assumption in most theoretical work showing closer convergence given more data. Another issue is that the 'ground truth labels' themselves often rely on shaky foundations, typically requiring human judgement. This applies, e.g., to multi-modal models which rely on human annotations. All these concerns also apply to simple token prediction underlying LLMs, as the 'ground truth' is defined by web-scraped text of varying quality, and the conditional distributions for the next token can hardly be said to be well-defined for a novel prompt.

4 Bengio et al. (2025) note this choice but do not discuss how it is made.

9.4 THE POWER OF NARRATIVE

> The only thing that makes life possible is permanent, intolerable uncertainty: not knowing what comes next.
>
> – Ursula K. Le Guin (1969): *The Left Hand of Darkness*

> In diversity is life and where there's life there's hope
>
> – Ursula K. Le Guin (1972): *The Word for World is Forest*

Given this shaky nature of the ground truths that AI systems as technical solutions are supposed to approximate, Leonelli (2025) rightfully emphasises the need for transdisciplinary approaches and domain knowledge. This is important for avoiding 'abstraction traps' (Selbst et al., 2019), as purely technical solutions with systems whose "'environment" is purely computational' (Bengio et al., 2025, p. 40) can only get us so far. However, Leonelli's vision of Environmental Intelligence does not stop at *acknowledging* these shaky ground truths and limitations of sterile AI systems; it also goes beyond more narrow technological or economic discussions and actually *embraces* these difficulties of modelling the world. As she argues in her third premise, the 'open-endedness' of human interactions with the world is actually desirable. While this notion remains a bit vague (in accordance with the concept itself), it fits seamlessly into her general narrative that celebrates the diversity and complexity of the natural and social world. It thus becomes clear that her vision fits much better to Ursula K. Le Guin's science fiction writing than to the popular stories she criticises.

As humans we like narratives, and the idea of human-like Artificial Intelligence that is either catastrophic or good for everyone is pervasive in public and academic imagination, not least due to its simplicity. According to the good-for-everyone version, the transcendence of human limitations will provide never-seen levels of productivity, leading to life improvements for everyone; this makes even work on resource-intensive AI capabilities seem beneficial by default. But as well documented by Acemoglu and Johnson (2023), powerful technologies do not come with an automatic productivity bandwagon effect that distributes the fruits of innovation among all parts of society. Not surprisingly, however, the narrative of benefiting 'all of humanity' is pushed by AI companies alongside the catastrophic version—allowing them to accumulate more publicity, investment, and deference—, while they internally define the achievement of 'Artificial General Intelligence' through the amount of generated corporate profit. Acemoglu and Johnson also highlight that often, small numbers of individuals and institutions are in a position to wield high power of persuasion regarding the development and adoption of technologies, which allows them

to have a large influence on important decisions for society as a whole. This involves the power of setting the agenda in discussions of technology and, in particular, of pushing narratives that further reinforce this concentration of power. In this way, the envisioned technological developments are presented as in the common interest and often as inevitable. Given these dynamics, it is crucial to provide alternative visions—not least for researchers and engineers working on the development of the technology of the future.

It is in this context that I see the significance of Leonelli's article. While it is not easy to provide a 'competitive' alternative, she connects a set of interrelated criticisms of the standard AI story, behind which a common narrative starts to emerge. Above, I elaborated on the rejection of solid ground truths and sterile notions of intelligence, which highlights that neither current systems nor future developments are automatic. Importantly, the ideal of sterile intelligence also reinforces the older trend to increasingly focus on more easily measurable quantities (Porter, 1995), thereby downplaying the importance or validity of complex (such as social or emotional) values or concepts. Indeed, 'one of the primary impulses of the modern state is to translate human experience into data readable by machines' (Recht, 2025b, p. 14). In a way, a focus on sterile intelligence can make the world itself more sterile. Another component EI in Leonelli's article is the recognition of the importance of our ecosystem on this planet (on both pragmatic and ethical grounds), which is all too often taken out of the equation. In line with Acemoglu and Johnson, she also criticises the concentration of power, lack of participation, and exacerbation of inequality. And while it is hardly possible to blend all these considerations into a single storyline, she does an admirable job in weaving the different themes into a cohesive narrative. She also derives practical suggestions, even if the path to getting them implemented is not fully visible yet. Here, a look into history may provide fruitful, leading Acemoglu and Johnson (2023) to call, for example, for more collective action across society. But also for such movements, 'a new narrative is critical' (p. 393); Environmental Intelligence agrees with many of their concerns and positive vision of shared prosperity, in a framing that more directly competes with the standard AI narrative. Leonelli contributes important foundations to this in her article, once again demonstrating that philosophy matters.

ACKNOWLEDGEMENTS

The author would like to thank Rabanus Derr for helpful feedback.

BH is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The author would like to thank

the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for their support.

EPILOGUE

The credibility of numbers [...] is a social and moral problem.
This has not yet been adequately appreciated.

— Theodore M. Porter, *Trust in Numbers*

We left the prologue with the question of whether predictive policing really makes predictions. My answer is that risk assessment always entails a prediction—and as such, it not only relies on modelling choices and inductive assumptions but should also be subjected to rigorous auditing. A probabilistic risk assessment involves prediction to the same extent as a weather forecast; a major difference is that, for the latter, we are not so easily misled into seeing the predicted risks (of, e.g., rain) as properties of the world—rather than of the model and past data. Another big difference is that weather forecasts are not performative; that is, the prediction does not (via a decision) influence the predicted outcome. This makes crime predictions particularly difficult to evaluate; this can be alleviated through causal modelling, in the sense of predicting what would happen under different interventions. More generally, the social world is less stable than many other domains that we model probabilistically, so the inductive assumptions we make here are in particular need of justification. It is equally important to carefully select and justify modelling choices, such as the attributes used to represent people—not only in terms of average performance (as is usually done), but also in terms of the broader social impacts. We have illustrated this concern by showing that using demographic attributes can exacerbate stereotypes or inequality.

Overall, it is important to *not* see algorithmic predictions as being on an inevitable march toward Minority Report–type oracles—which aren't even reliable in the story, for performativity reasons—but as a narrow tool that can predict well *on average* under certain stability assumptions. Furthermore, statistical tools only offer a particular and limited approach to addressing problems like crime or allocation, and should always be considered alongside alternative strategies. As put by Ben Recht (2025b, p. 14), opting for statistical tools 'narrows the possibilities of the world we inhabit' as it 'demands machine-readable rules, data, and outcomes'. In line with my earlier assistance on thorough justification of modelling assumptions and choices, he emphasises that 'if we're going to

rely on statistical prediction, then we need expertise in statistical hygiene'. This echoes quite closely a warning from 101 years earlier, that 'it will probably always be found that those statisticians who use mathematical formulas without guarding well their limitations are likely to have influence productive of an evil very similar to that produced by Quetelet' (Rietz, 1924, p. 418). While we have made progress since Rietz's days, our fundamental understanding of statistical methods is still in need of improvement. In particular, the 'issues in the foundations of statistics' that David Freedman (1995, p. 35) identified still loom large: 'How do statistical models connect with reality? What areas lend themselves to investigation by statistical modelling? When are such investigations likely to be sterile?' Much of this thesis can be read as a response to the first question and a motivation for the other two. Not surprisingly, then, my recommendations for future research have substantial overlap with Freedman's questions.[5]

To further investigate notions of stability, it would be important to extend the proposed modelling of finite populations and, in particular, connect it to existing notions of distribution shift and robust optimisation. A complementary direction is to investigate to what extent relevant assumptions tend to be satisfied in different social domains; this requires a large-scale empirical investigation using suitable datasets. Another avenue is to work on a more tangible understanding of the effects of data collection choices, both in terms of attributes and in terms of the considered population. I have considered demographic attributes in this thesis, but this analysis can be extended and related to existing work in the social sciences. It would also be important to account more for the lived reality of affected populations in model construction and auditing, starting by creating formats for conversations between researchers and communities.[6]

Overall, in order to better anticipate societal consequences and to set standards for accountability, it is necessary to establish best practices for context-sensitive justification and validation of modelling choices and assumptions.

5 In general, Freedman is perhaps the author with whom I have found myself agreeing the most. The works with the strongest influence on the path I took during my PhD have probably been Porter (1995), see epigraph, and Moss (2021), documenting this phenomenon in machine learning.

6 Building on our position paper, we are taking a step in this direction by organising a workshop at CHI 2026: https://sites.google.com/view/between-and-beyond.

BIBLIOGRAPHY

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge (2014). *Finite population causal standard errors*. Tech. rep. National Bureau of Economic Research Cambridge.

— (2020). 'Sampling-based versus design-based uncertainty in regression analysis'. In: *Econometrica* 88.1, pp. 265–296.

Abdu, Amina A, Irene V Pasquetto and Abigail Z Jacobs (2023). 'An empirical analysis of racial categories in the algorithmic fairness literature'. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1324–1333.

Abebe, Rediet, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan and David G Robinson (2020). 'Roles for computing in social change'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 252–260.

Acemoglu, Daron and Simon Johnson (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Basic Books UK.

Acerbi, Carlo (2002). 'Spectral measures of risk: A coherent representation of subjective risk aversion'. In: *Journal of Banking & Finance* 26.7, pp. 1505–1518.

Ajmal, Tananant Boonya-Ananta, Andres J Rodriguez, VN Du Le and Jessica C Ramella-Roman (2021). 'Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal'. In: *Biomedical Optics Express* 12.12, pp. 7445–7457.

Alghamdi, Wael, Hsiang Hsu, Haewon Jeong, Hao Wang, P Winston Michalak, Shahab Asoodeh and Flavio P Calmon (2022). 'Beyond Adult and COMPAS: Fairness in multi-class prediction'. In: *arXiv preprint arXiv:2206.07801*.

Ali, Suki (2020). *Mixed-Race, Post-Race: Gender, New Ethnicities and Cultural Practices*. Routledge.

Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill and Astrid Mager (2020). 'Algorithmic profiling of job seekers in Austria: How austerity politics are made effective'. In: *Frontiers in Big Data*, p. 5.

Alt, Franz L. (1972). 'Archaelogy of computers: Reminiscences, 1945-1947'. In: *Communications of the ACM* 15.7, pp. 693–694.

American Anthropological Association (1998). 'AAA Statement on Race'. In: *Anthropology Newsletter* 39.9, p. 3.

Andrés-Pueyo, Antonio, Karin Arbach-Lucioni and Santiago Redondo (2018). 'The RisCanvi: a new tool for assessing risk for violence in prison and recidivism'. In: *Handbook of Recidivism Risk/needs Assessment Tools*, pp. 255–268.

Angrist, Joshua D (1990). 'Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records'. In: *The American Economic Review*, pp. 313–336.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin (1996). 'Identification of causal effects using instrumental variables'. In: *Journal of the American statistical Association* 91.434, pp. 444–455.

Angrist, Joshua D and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

— (2010). 'The credibility revolution in empirical economics: How better research design is taking the con out of econometrics'. In: *Journal of Economic Perspectives* 24.2, pp. 3–30.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner (2016). 'Machine Bias'. In: *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Appiah, Anthony (2017). *As If: Idealization and Ideals*. Harvard University Press.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber and David Heath (1999). 'Coherent measures of risk'. In: *Mathematical Finance* 9.3, pp. 203–228.

Athey, Susan (2017). 'Beyond prediction: Using big data for policy problems'. In: *Science* 355.6324, pp. 483–485.

Athey, Susan and Guido W Imbens (2016). 'Recursive partitioning for heterogeneous causal effects'. In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Bach, Ruben L, Christoph Kern, Hannah Mautner and Frauke Kreuter (2023). 'The impact of modeling decisions in statistical profiling'. In: *Data & Policy* 5, e32.

Ball, Elena, Melanie C Steffens and Claudia Niedlich (2022). 'Racism in Europe: characteristics and intersections with other social categories'. In: *Frontiers in Psychology* 13, p. 789661.

Bardenet, Rémi and Odalric-Ambrym Maillard (2015). 'Concentration inequalities for sampling without replacement'. In: *Bernoulli* 21.3, pp. 1361–1385.

Bareinboim, Elias and Judea Pearl (2013). 'A general algorithm for deciding transportability of experimental results'. In: *Journal of Causal Inference* 1.1, pp. 107–134.

Barnes, Sally-Anne, Sally Wright, Pat Irving and Isabelle Deganis (2015). *Identification of Latest Trends and Current Developments in Methods to Profile Jobseekers in European Public Employment Services: Final Report*. Tech. rep. Brussels: European Commission, Directorate-General for Employment, Social Affairs and Inclusion. URL: https://ec.europa.eu/social/BlobServlet?docId=14173\&langId=en.

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

Barocas, Solon and Andrew D Selbst (2016). 'Big data's disparate impact'. In: *California Law Review* 104, p. 671.

Beck, Ulrich (1986). 'Risikogesellschaft: Auf dem Weg in eine andere Moderne'. In.

Becker, Ann-Kristin, Oana Dumitrasc and Klaus Broelemann (2024). 'Standardized Interpretable Fairness Measures for Continuous Risk Scores'. In: *International Conference on Machine Learning*.

Becker, Barry and Ronny Kohavi (1996). *Adult*. UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/dataset/2/adult.

Belitz, Clara, Jaclyn Ocumpaugh, Steven Ritter, Ryan S Baker, Stephen E Fancsali and Nigel Bosch (2023). 'Constructing categories: Moving beyond protected classes in algorithmic fairness'. In: *Journal of the Association for Information Science and Technology* 74.6, pp. 663–668.

Bell, Eric Temple (1945). *The Development of Mathematics*. 2nd ed. McGraw-Hill Book Company.

Bengio, Yoshua, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson et al. (2025). 'Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path?' In: *arXiv preprint arXiv:2502.15657*.

Benjamin, Ruha (2019). *Race After Technology: Abolitionist Tools for the New Jim Cde*. John Wiley & Sons.

Benthall, Sebastian and Bruce D Haynes (2019). 'Racial categories in machine learning'. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 289–298.

Bereni, Laure, Renaud Epstein and Manon Torres (2021). 'Colour-blind diversity: how the | "Diversity Label" reshaped anti-discrimination policies in three French local governments'. In: *Diversity in Local Political Practice*. Routledge, pp. 14–32.

Berk, Richard A (1987). 'Causal inference as a prediction problem'. In: *Crime and Justice* 9, pp. 183–200.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth (2021). 'Fairness in criminal justice risk assessments: The state of the art'. In: *Sociological Methods & Research* 50.1, pp. 3–44.

Berman, Alexander, Karl de Fine Licht and Vanja Carlsson (2024). 'Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system'. In: *Technology in Society* 76, p. 102471.

Bernoulli, Jakob (1713). *Ars Coniectandi*. Impensis Thurnisiorum, fratrum.

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou and Aylin Caliskan (2023). 'Easily accessible text-to-image generation amplifies demographic stereotypes at large scale'. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel and Shakir Mohamed (2022a). 'Power to the people? Opportunities and challenges for participatory AI'. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–8.

Birhane, Abeba and Marek McGann (2024). 'Large models of what? Mistaking engineering achievements for human linguistic agency'. In: *Language Sciences* 106, p. 101672.

Birhane, Abeba, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque and Christopher L. Dancy (2022b). 'The forgotten margins of AI ethics'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 948–958.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Springer.

Black, Emily, John Logan Koepke, Pauline Kim, Solon Barocas and Mingwei Hsu (2024a). 'Less discriminatory algorithms'. In: *Georgetown Law Journal* 113.1.

Black, Emily, Logan Koepke, Pauline Kim, Solon Barocas and Mingwei Hsu (2024b). 'The legal duty to search for less discriminatory algorithms'. In: *arXiv preprint arXiv:2406.06817*.

Black, Emily, Manish Raghavan and Solon Barocas (2022). 'Model multiplicity: Opportunities, concerns, and solutions'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863.

Bloomberg (2023). *Generative AI: What Happens When Bias Is Built In?* `https://www.bloomberg.com/graphics/2023-generative-ai-bias/?utm_source=website&utm_medium=share&utm_campaign=copy`. Accessed: 2025-01-13.

Bo, Hao and Sebastian Galiani (2021). 'Assessing external validity'. In: *Research in Economics* 75.3, pp. 274–285.

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain and Lucy Vasserman (2019). 'Nuanced metrics for measuring unintended bias with real data for text classification'. In: *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500.

Bovens, Mark (2010). 'Two concepts of accountability: Accountability as a virtue and as a mechanism'. In: *West European Politics* 33.5, pp. 946–967.

Bowker, Geoffrey and Susan Leigh Star (1999). *Sorting Things Out: Classification and Its Consequences*. Vol. 4. MIT Press.

Boyd, Danah and Kate Crawford (2012). 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon'. In: *Information, Communication & Society* 15.5, pp. 662–679.

Braveman, Paula and Tyan Parker Dominguez (2021). 'Abandon "race." Focus on racism'. In: *Frontiers in Public Health* 9, p. 689462.

Breiman, Leo (2001). 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)'. In: *Statistical Science* 16.3, pp. 199–231.

Brier, Glenn W (1950). 'Verification of forecasts expressed in terms of probability'. In: *Monthly Weather Review* 78.1, pp. 1–3.

Brocket, Tom (2020). 'From "in-betweenness" to "positioned belongings": Second-generation Palestinian-Americans negotiate the tensions of assimilation and transnationalism'. In: *Ethnic and Racial Studies* 43.16, pp. 135–154.

Brown Jr, Terry D, Francis C Dane and Marcus D Durham (1998). 'Perception of race and ethnicity.' In: *Journal of Social Behavior & Personality* 13.2.

Buchak, Lara (2013). *Risk and Rationality*. Oxford University Press.

Buolamwini, Joy and Timnit Gebru (2018). 'Gender shades: Intersectional accuracy disparities in commercial gender classification'. In: *2018 Conference on Fairness, Accountability, and Transparency*. PMLR, pp. 77–91.

Burgund, E Darcy, Yiyang Zhao, Inaya N Laubach and Eyerusalem F Abebaw (2024). 'Different features for different races: Tracking the eyes of Asian, Black, and White participants viewing Asian, Black, and White Faces'. In: *PLoS One* 19.9, e0310638.

Burhanpurkar, Maya, Zhun Deng, Cynthia Dwork and Linjun Zhang (2021). 'Scaffolding sets'. In: *arXiv preprint arXiv:2111.03135*.

Butler, Judith (2011). *Bodies That Matter: On the Discursive Limits of Sex*. Routledge.

CJEU (2017). *Judgment of 6 April 2017, Jyske Finans A/s v. Ligebehandlingsnævnet*. C-668/15 ECLI:EU:C:2017:278.

Cabitza, Federico, Andrea Campagner and Lorenzo Famiglini (2022). 'Global interpretable calibration index, a new metric to estimate machine learning

models' calibration'. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, pp. 82–99.

Calders, Toon and Sicco Verwer (2010). 'Three naive Bayes approaches for discrimination-free classification'. In: *Data Mining and Knowledge Discovery* 21, pp. 277–292.

Camara, Nina (2016). *Lost in Otherness: Growing Up as a Mixed-Raced Child in Eastern Europe*. Accessed: 2025-01-02. URL: https://afropean.com/lost-in-otherness-growing-up-as-a-mixed-raced-child-in-eastern-europe/.

Cambridge Dictionary (2025). *Reification - Definition in the Cambridge English Dictionary*. Accessed: 20-Jan-2025. URL: https://dictionary.cambridge.org/dictionary/english/reification.

Campbell, Donald T (1958). 'Common fate, similarity, and other indices of the status of aggregates of persons as social entities'. In: *Behavioral Science* 3.1, p. 14.

Cao, Qiong, Li Shen, Weidi Xie, Omkar M Parkhi and Andrew Zisserman (2018). 'Vggface2: A dataset for recognising faces across pose and age'. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 67–74.

Card, David and Alan B Krueger (1994). 'Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania'. In: *The American Economic Review* 84.4, p. 772.

Carnap, Rudolf (1950). *Logical Foundations of Probability*. University of Chicago Press.

Cartwright, Nancy (1999). 'The limits of exact science, from economics to physics'. In: *Perspectives on Science* 7.3, pp. 318–336.

Chekroud, Adam M, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R Corlett, Nikolaos Koutsouleris, Harlan M Krumholz et al. (2024). 'Illusory generalizability of clinical prediction models'. In: *Science* 383.6679, pp. 164–167.

Chen, Daniel (2020). *FairFace*. https://github.com/dchen236/FairFace. Accessed: 2025-01-15.

Chen, Jiahao, Nathan Kallus, Xiaojie Mao, Geoffry Svacha and Madeleine Udell (2019). 'Fairness under unawareness: Assessing disparity when protected class is unobserved'. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 339–348.

Chen, Tianwei, Yusuke Hirota, Mayu Otani, Noa Garcia and Yuta Nakashima (2024). 'Would deep generative models amplify bias in future models?' In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10833–10843.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins (2018). 'Double/debiased machine learning for treatment and structural parameters'. In: *The Econometrics Journal* 21.1, pp. C1–C68.

Chouldechova, Alexandra (2017). 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments'. In: *Big Data* 5.2, pp. 153–163.

Cikara, Mina, Joel E Martinez and Neil A Lewis Jr (2022). 'Moving beyond social categories by incorporating context in social psychological theory'. In: *Nature Reviews Psychology* 1.9, pp. 537–549.

Cobbe, Jennifer, Michelle Seng Ah Lee and Jatinder Singh (2021). 'Reviewable automated decision-making: A framework for accountable algorithmic systems'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 598–609.

Commission, European (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. Accessed: 2024-01-24. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Commission, European, Directorate-General for Justice, Consumers, European network of legal experts in gender equality, non discrimination, I. Chopin and C. Germaine (2024). *A comparative analysis of non-discrimination law in Europe 2023 – The 27 EU Member States compared*. Publications Office of the European Union. DOI: doi/10.2838/990418.

Coolidge, Julian Lowell (1921). 'The dispersion of observations'. In: *Bulletin of the American Mathematical Society* 27.9-10, pp. 439–442.

Cooper, A Feder, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen and Baobao Zhang (2024). 'Arbitrariness and social prediction: The confounding role of variance in fair classification'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 20, pp. 22004–22012.

Coots, Madison, Soroush Saghafian, David Kent and Sharad Goel (2023). 'Reevaluating the role of race and ethnicity in diabetes screening'. In: *arXiv preprint arXiv:2306.10220*.

Corbett-Davies, Sam, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff and Sharad Goel (2023). 'The measure and mismeasure of fairness'. In: *The Journal of Machine Learning Research* 24.1, pp. 14730–14846.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel and Aziz Huq (2017). 'Algorithmic decision making and the cost of fairness'. In: *Proceed-*

ings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806.

Corbett, Eric, Emily Denton and Sheena Erete (2023). 'Power and public participation in ai'. In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13.

Corey, Ethan (2019). 'How a tool to help judges may be leading them astray'. In: *The Appeal, August* 8. Accessed: 2025-12-11.

Coston, Amanda, Ashesh Rambachan and Alexandra Chouldechova (2021). 'Characterizing fairness over the set of good models under selective labels'. In: *International Conference on Machine Learning*. PMLR, pp. 2144–2155.

Cournot, Antoine Augustin (1843). *Exposition de la théorie des chances et des probabilités*. L. Hachette.

Crawford, Kate and Trevor Paglen (2021). 'Excavating AI: The politics of images in machine learning training sets'. In: *AI & Society* 36.4, pp. 1105–1116.

Crisan, Anamaria, Margaret Drouhard, Jesse Vig and Nazneen Rajani (2022). 'Interactive model cards: A human-centered approach to model documentation'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 427–439.

Cruz, André and Moritz Hardt (2024). 'Unprocessing Seven Years of Algorithmic Fairness'. In: *The Twelfth International Conference on Learning Representations*.

Currie, Geoffrey, Johnathan Hewis, Elizabeth Hawk and Eric Rohren (2024). 'Gender and ethnicity bias of text-to-image generative artificial intelligence in medical imaging, part 1: preliminary evaluation'. In: *Journal of Nuclear Medicine Technology* 52.4, pp. 356–359.

D'Incà, Moreno, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi and Nicu Sebe (2024). 'OpenBias: Open-set bias detection in text-to-image generative models'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12225–12235.

Danks, David (2015). 'Goal-dependence in (scientific) ontology'. In: *Synthese* 192, pp. 3601–3616.

Dawid, A Philip (1982). 'The well-calibrated Bayesian'. In: *Journal of the American Statistical Association* 77.379, pp. 605–610.

— (1985). 'Calibration-based empirical probability'. In: *The Annals of Statistics* 13.4, pp. 1251–1274.

Dawid, Philip (2000). 'Causal inference without counterfactuals'. In: *Journal of the American statistical Association* 95.450, pp. 407–424.

— (2017). 'On individual risk'. In: *Synthese* 194.9, pp. 3445–3474.

— (2021). 'Decision-theoretic foundations for statistical causality'. In: *Journal of Causal Inference* 9.1, pp. 39–77.

— (2022). 'Decision-theoretic foundations for statistical causality: Response to Pearl'. In: *Journal of Causal Inference* 10.1, pp. 296–299.

De Cooman, Gert and Jasper De Bock (2022). 'Randomness is inherently imprecise'. In: *International Journal of Approximate Reasoning* 141, pp. 28–68.

De Finetti, Bruno (1937). 'La prévision: ses lois logiques, ses sources subjectives'. In: *Annales de l'institut Henri Poincaré*. Vol. 7. 1, pp. 1–68.

De Finetti, Bruno and Leonard J Savage (1962). 'Sul modo di scegliere le probabilità iniziali'. In: *Biblioteca del Metron, Serie C* 1, pp. 81–154.

DeGroot, Morris H and Stephen E Fienberg (1983). 'The comparison and evaluation of forecasters'. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 32.1-2, pp. 12–22.

Deaton, Angus (2009). *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*. Tech. rep. National Bureau of Economic Research.

Deaton, Angus and Nancy Cartwright (2018). 'Understanding and misunderstanding randomized controlled trials'. In: *Social Science & Medicine* 210, pp. 2–21.

Delgado, Fernando, Solon Barocas and Karen Levy (2022). 'An uncommon task: Participatory design in legal AI'. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1, pp. 1–23.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei (2009). 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, pp. 248–255.

Derr, Rabanus, Jessie Finocchiaro and Robert C Williamson (2025). 'Three types of calibration with properties and their semantic and formal relationships'. In: *arXiv preprint arXiv:2504.18395*.

Derr, Rabanus and Robert C Williamson (2023). 'Systems of precision: Coherent probabilities on pre-Dynkin systems and coherent previsions on linear subspaces'. In: *Entropy* 25.9, p. 1283.

— (2024). 'Fairness and randomness in machine learning: Statistical independence and relativization'. In: *The New England Journal of Statistics in Data Science* 3.1, pp. 55–72.

Desiere, Sam, Kristine Langenbucher and Ludo Struyven (2019). 'Statistical profiling in public employment services. An international comparison'. In: *OECD Social, Employment and Migration Working Papers* 224.

Desrosières, Alain (1997). 'Quetelet et la sociologie quantitative: du piédestal à l'oubli (Quetelet and quantitative sociology: from the pedestal to oblivion)'.

In: *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet. Actes du colloque des*. Vol. 24, pp. 179–198.

Detlef, Detlef Dürr and Stefan Teufel (2009). *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer.

Devroye, Luc, Laszlo Gyorfi, Adam Krzyzak and Gábor Lugosi (1994). 'On the strong universal consistency of nearest neighbor regression function estimates'. In: *The Annals of Statistics* 22.3, pp. 1371–1385.

Ding, Frances, Moritz Hardt, John Miller and Ludwig Schmidt (2021). 'Retiring adult: New datasets for fair machine learning'. In: *Advances in Neural Information Processing Systems* 34, pp. 6478–6490.

Doh, Miriam, Benedikt Höltgen, Piera Riccio and Nuria Oliver (2025). 'Position: The categorization of race in ML is a flawed premise'. In: *International Conference on Machine Learning*.

Donoho, David (2024). 'Data science at the singularity'. In: *Harvard Data Science Review* 6.1. https://hdsr.mitpress.mit.edu/pub/g9mau4mo.

Douglas, Heather (2000). 'Inductive risk and values in science'. In: *Philosophy of Science* 67.4, pp. 559–579.

Douglas, Mary (1992). *Risk and Blame: Essays in Cultural Theory*. Routledge.

Dwork, Cynthia (2022). 'Fairness, randomness, and the crystal ball'. In: *Munich AI Lectures*. URL: https://www.youtube.com/watch?v=n4XftI9G0fA (visited on 08/04/2023).

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel (2012). 'Fairness through awareness'. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226.

Dwork, Cynthia, Chris Hays, Nicole Immorlica, Juan C Perdomo and Pranay Tankala (2025). 'From fairness to infinity: Outcome-indistinguishable (omni) prediction in evolving graphs'. In: *The Thirty Eighth Annual Conference on Learning Theory*.

Egami, Naoki and Erin Hartman (2023). 'Elements of external validity: Framework, design, and analysis'. In: *American Political Science Review* 117.3, pp. 1070–1088.

Egbert, Simon, Vasilis Galis, Helene Oppen Ingebrigtsen Gundhus and Christin Thea Wathne (2024). 'The platformization of policing: A cross-national analysis'. In: *Policing and Intelligence in the Global Big Data Era, Volume I: New Global Perspectives on Algorithmic Governance*. Springer, pp. 349–392.

Eide, Elisabeth (2016). 'Strategic essentialism'. In: *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pp. 2278–2280.

Eliot, George (1871). *Middlemarch*. William Blackwood and Sons.

Elliott, Michael R and Richard Valliant (2017). 'Inference for nonprobability samples'. In: *Statistical Science* 32.2, pp. 249–264.

Engler, Alex (2023). *The EU and US diverge on AI regulation: A transatlantic comparison and steps to alignment*. Accessed: 24-01-2025. URL: https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/.

Equivant (formerly Northpointe, Inc.) (2017). *Practitioner's Guide to COMPAS Core*. Technical Guide / Practitioner Manual. Accessed: 2025-12-11. Equivant. URL: https://cjdata.tooltrack.org/sites/default/files/2018-10/Practitioners_Guide_COMPASCore_121917.pdf.

European Commission (2017). *Data Collection in the Field of Ethnicity*. Accessed: 2025-01-09. URL: https://commission.europa.eu/system/files/2021-09/data_collection_in_the_field_of_ethnicity.pdf.

Fakhro, Abdulla, Hyung Woo Yim, Yong Kyu Kim and Anh H Nguyen (2015). 'The evolution of looks and expectations of Asian eyelid and eye appearance'. In: *Seminars in plastic surgery*. Vol. 29. 03. Thieme Medical Publishers, pp. 135–144.

Fazelpour, Sina and Zachary C Lipton (2020). 'Algorithmic fairness from a non-ideal perspective'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 57–63.

Feduzi, Alberto, Jochen Runde and Carlo Zappia (2012). 'De Finetti on the insurance of risks and uncertainties'. In: *The British Journal for the Philosophy of Science*.

Fekete, Liz (2014). 'Europe against the Roma'. In: *Race & Class* 55.3, pp. 60–70.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger and Suresh Venkatasubramanian (2015). 'Certifying and removing disparate impact'. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.

Feliciano, Cynthia (2016). 'Shades of race: How phenotype and observer characteristics shape racial classification'. In: *American Behavioral Scientist* 60.4, pp. 390–419.

Feng, Yiding and Wei Tang (2025). 'Measuring informativeness gap of (mis) calibrated predictors'. In: *arXiv preprint arXiv:2507.12094*.

Findley, Michael G, Kyosuke Kikuta and Michael Denly (2021). 'External validity'. In: *Annual Review of Political Science* 24.1, pp. 365–393.

Fisher, Ronald A (1922). 'On the mathematical foundations of theoretical statistics'. In: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604, pp. 309–368.

Fisher, Ronald Aylmer (1925). 'Statistical methods for research workers'. In.

Fisher, Ronald (1955). 'Statistical methods and scientific induction'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 17.1, pp. 69–78.

Floro, Marissa (2018). 'In between: What the experiences of biracial, bisexual women tell us about identity formation'. PhD thesis. Loyola University Chicago.

Flyvbjerg, Bent, Carsten Glenting and Arne Rønnest (2004). 'Procedures for dealing with optimism bias in transport planning'. In: *London: The British Department for Transport, Guidance Document*.

Ford, Karly S, Ashley N Patterson and Marc P Johnston-Guerrero (2021). 'Monoracial normativity in university websites: Systematic erasure and selective reclassification of multiracial students.' In: *Journal of Diversity in Higher Education* 14.2, p. 252.

Foster, Dean P and Sergiu Hart (2021). 'Forecast hedging and calibration'. In: *Journal of Political Economy* 129.12, pp. 3447–3490.

Foster, Dean P and Rakesh V Vohra (1998). 'Asymptotic calibration'. In: *Biometrika* 85.2, pp. 379–390.

Fox, Matthew P, Eleanor J Murray, Catherine R Lesko and Shawnita Sealy-Jefferson (2022). 'On the need to revitalize descriptive epidemiology'. In: *American Journal of Epidemiology* 191.7, pp. 1174–1179.

Freedman, David A (1995). 'Some issues in the foundation of statistics'. In: *Topics in the Foundation of Statistics*, pp. 19–39.

— (2006). 'Statistical models for causation: What inferential leverage do they provide?' In: *Evaluation Review* 30.6, pp. 691–713.

— (2008). 'On regression adjustments to experimental data'. In: *Advances in Applied Mathematics* 40.2, pp. 180–193.

Freedman, David A and Richard A Berk (2008). 'Weighting regressions by propensity scores'. In: *Evaluation Review* 32.4, pp. 392–409.

Freedman, David (1997). 'Some issues in the foundation of statistics'. In: *Topics in the Foundation of Statistics*, pp. 19–39.

Frisch, R (1930). *A Dynamic Approach to Economic Theory: Lectures by Ragnar Frish at Yale University*.

Fröhlich, Christian and Robert C Williamson (2024a). 'Insights from insurance for fair machine learning: Responsibility, performativity and aggregates'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

— (2024b). 'Risk measures and upper probabilities: Coherence and stratification'. In: *Journal of Machine Learning Research* 25.207, pp. 1–100.

Fu, Siyao, Haibo He and Zeng-Guang Hou (2014). 'Learning race from face: A survey'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.12, pp. 2483–2509.

Gamper, Jutta, Günter Kernbeiß and Michael Wagner-Pinter (2020). *Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung*.

Garber, Daniel and Sandy Zabell (1979). 'On the emergence of probability'. In: *Archive for History of Exact Sciences*, pp. 33–53.

Garcia, Noa, Yusuke Hirota, Yankun Wu and Yuta Nakashima (2023). 'Uncurated image-text datasets: Shedding light on demographic bias'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6957–6966.

Garcia, Raul Vicente, Lukasz Wandzik, Louisa Grabner and Joerg Krueger (2019). 'The harms of demographic bias in deep face recognition research'. In: *2019 International Conference on Biometrics (ICB)*. IEEE, pp. 1–6.

Garland, David (2001). *The Culture of Control: Crime and Social Order in Contemporary Society*. Oxford University Press.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (2021). 'Datasheets for datasets'. In: *Communications of the ACM* 64.12, pp. 86–92.

Georgopoulos, Markos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis and Maja Pantic (2021). 'Mitigating demographic bias in facial datasets with style-based multi-attribute transfer'. In: *International Journal of Computer Vision* 129.7, pp. 2288–2307.

Ghosh, Sourojit, Nina Lutz and Aylin Caliskan (2024). '"I don't see myself represented here at all": User experiences of stable diffusion outputs containing representational harms across gender identities and nationalities'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7, pp. 463–475.

Gigerenzer, Gerd (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.

Gigerenzer, Gerd, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo and Konstantinos V Katsikopoulos (2005). '"A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts?' In: *Risk Analysis: An International Journal* 25.3, pp. 623–629.

Gillis, Talia B, Vitaly Meursault and Berk Ustun (2024). 'Operationalizing the search for less discriminatory alternatives in fair lending'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 377–387.

Giuliani, Luca, Eleonora Misino and Michele Lombardi (2023). 'Generalized disparate impact for configurable fairness solutions in ML'. In: *International Conference on Machine Learning*. PMLR, pp. 11443–11458.

Glasgow, Russell E, Lawrence W Green, Lisa M Klesges, David B Abrams, Edwin B Fisher, Michael G Goldstein, Laura L Hayman, Judith K Ockene and C Tracy Orleans (2006). 'External validity: We need to do more.' In: *Annals of Behavioral Medicine* 31.2.

Glymour, Clark (2001). 'Instrumental probability'. In: *The Monist* 84.2, pp. 284–300.

Gneiting, Tilmann and Adrian E Raftery (2007). 'Strictly proper scoring rules, prediction, and estimation'. In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodman, Nelson (1972). 'Seven strictures on similarity'. In: *Problems and Projects*. Indianapolis: Bobbs-Merrill, pp. 437–447.

Google AI (2024). *Gemini 2.0 Flash*. https://ai.google.dev/gemini-api/docs/models/gemini-v2. Accessed: 2025-01-13.

Gopalan, Parikshit, Michael P Kim, Mihir A Singhal and Shengjia Zhao (2022). 'Low-degree multicalibration'. In: *The Thirty Fifth Annual Conference on Learning Theory*. PMLR, pp. 3193–3234.

Gopalan, Parikshit, Michael Kim and Omer Reingold (2023). 'Swap agnostic learning, or characterizing omniprediction via multicalibration'. In: *Advances in Neural Information Processing Systems* 36, pp. 39936–39956.

Grabisch, Michel, Jean-Luc Marichal, Radko Mesiar and Endre Pap (2009). *Aggregation Functions*. Cambridge University Press.

Greblicki, Włodzimierz, Adam Krzyżak and Mirosław Pawlak (1984). 'Distribution-free pointwise consistency of kernel regression estimate'. In: *The Annals of Statistics* 12.4, pp. 1570–1575.

Green, Ben (2022). 'Escaping the impossibility of fairness: From formal to substantive algorithmic fairness'. In: *Philosophy & Technology* 35.4, p. 90.

Greenland, Sander (2012). 'Causal inference as a prediction problem: Assumptions, identification and evidence synthesis'. In: *Causality: Statistical Perspectives and Applications*, pp. 43–58.

— (2017). 'For and against methodologies: Some perspectives on recent causal and statistical inference debates'. In: *European Journal of Epidemiology* 32, pp. 3–20.

Groh, Matthew, Caleb Harris, Roxana Daneshjou, Omar Badri and Arash Koochek (2022). 'Towards transparency in dermatology image datasets

with skin tone annotations by experts, crowds, and an algorithm'. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2, pp. 1–26.

Guo, Chuan, Geoff Pleiss, Yu Sun and Kilian Q Weinberger (2017). 'On calibration of modern neural networks'. In: *International Conference on Machine Learning*. PMLR, pp. 1321–1330.

Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi et al. (2025). 'Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning'. In: *arXiv preprint arXiv:2501.12948*.

Hacking, Ian (1975). *The Emergence of Probability*. Cambridge University Press.

— (1986). 'Making up people'. In: *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*, pp. 222–236.

— (1990). *The Taming of Chance*. Cambridge University Press.

Hájek, Alan (1996). '"Mises redux"–redux: Fifteen arguments against finite frequentism'. In: *Erkenntnis* 45, pp. 209–227.

— (2007). 'The reference class problem is your problem too'. In: *Synthese* 156, pp. 563–585.

— (2009). 'Fifteen arguments against hypothetical frequentism'. In: *Erkenntnis* 70.2, pp. 211–235.

— (2019). 'Interpretations of probability'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.

Hangartner, Dominik, Daniel Kopp and Michael Siegenthaler (2021). 'Monitoring hiring discrimination through online recruitment platforms'. In: *Nature* 589.7843, pp. 572–576.

Hanna, Alex, Emily Denton, Andrew Smart and Jamila Smith-Loud (2020). 'Towards a critical race methodology in algorithmic fairness'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512.

Hardt, Moritz (2014). *How big data is unfair: Understanding unintended sources of unfairness in data driven decision making*. Accessed: 2025-12-15. Medium. URL: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

— (2025). *The Emerging Science of Machine Learning Benchmarks*. Online at https://mlbenchmarks.org. Manuscript.

Hardt, Moritz, Eric Price and Nati Srebro (2016). 'Equality of opportunity in supervised learning'. In: *Advances in Neural Information Processing Systems* 29.

Harvey Adam. LaPlace, Jules. (2021). *Exposing.ai*. URL: https://exposing.ai (visited on 01/01/2021).

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

Heckman, James J and Rodrigo Pinto (2024). 'Econometric causality: The central role of thought experiments'. In: *Journal of Econometrics* 243.1-2, p. 105719.

Heckman, James J and Sergio Urzua (2010). 'Comparing IV with structural models: What simple IV can and cannot identify'. In: *Journal of Econometrics* 156.1, pp. 27–37.

Hedden, Brian (2013). 'Incoherence without exploitability'. In: *Noûs* 47.3, pp. 482–495.

Heljakka, Ari, Martin Trapp, Juho Kannala and Arno Solin (2022). 'Disentangling model multiplicity in deep learning'. In: *arXiv preprint arXiv:2206.08890*.

Hernán, Miguel A (2016). 'Does water kill? A call for less casual causal inferences'. In: *Annals of Epidemiology* 26.10, pp. 674–680.

Holl, Jürgen, Günter Kernbeiß and Michael Wagner-Pinter (2018). 'Das AMS-Arbeitsmarktchancen-Modell'. In: *Arbeitsmarktservice Österreich, Wien*.

Holland, Paul W (1986). 'Statistics and causal inference'. In: *Journal of the American statistical Association* 81.396, pp. 945–960.

Höltgen, Benedikt (2025). 'A thoughtful narrative emerges'. In: *Harvard Data Science Review* 7.4.

— (2026). 'A unifying perspective on probabilities as model predictions'. Preprint.

Höltgen, Benedikt and Nuria Oliver (2025). 'Reconsidering fairness through unawareness from the perspective of model multiplicity'. In: *EAAMO '25: Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

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

— (2026a). 'The accountability gap in causal modelling for automated decision systems'. Preprint.

— (2026b). 'The costs of pretending that there are data-generating probability distributions in the social world'. In: *FAccT '26: Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*.

Hong, Sun-ha (2023). 'Prediction as extraction of discretion'. In: *Big Data & Society* 10.1, p. 20539517231171053.

Horvitz, Daniel G and Donovan J Thompson (1952). 'A generalization of sampling without replacement from a finite universe'. In: *Journal of the American statistical Association* 47.260, pp. 663–685.

Hu, Lily (2023). 'Causal inference and the problem of variable choice'. In: *Social Foundations for Statistics and Machine Learning Workshop*. URL: https://www.youtube.com/watch?v=zSwAVwypMSs (visited on 03/01/2024).

Hu, Lily and Issa Kohler-Hausmann (2020). 'What's sex got to do with machine learning?' In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513.

— (2024). 'What is perceived when race is perceived and why it matters for causal inference and discrimination studies'. In: *Law & Society Review*, pp. 1–26.

Huang, Linzhi, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang and Jian Zhao (2023). 'Gradient attention balance network: Mitigating face recognition racial bias via gradient attention'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38–47.

Huggingface (2024). *Stable Diffusion 3.5 Large*. Accessed: 2025-01-26. URL: https://huggingface.co/stabilityai/stable-diffusion-3.5-large.

Hume, David (1777). *An Enquiry Concerning Human Understanding*.

Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes and Margaret Mitchell (2021). 'Towards accountability for machine learning datasets: Practices from software engineering and infrastructure'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 560–575.

Iacus, Stefano M, Gary King and Giuseppe Porro (2012). 'Causal inference without balance checking: Coarsened exact matching'. In: *Political Analysis* 20.1, pp. 1–24.

Imbens, Guido W (2020). 'Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics'. In: *Journal of Economic Literature* 58.4, pp. 1129–1179.

Imbens, Guido W and Joshua D Angrist (1994). 'Identification and estimation of local average treatment effects'. In: *Econometrica* 62.2, pp. 467–475.

Imbens, Guido W and Donald B Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, Guido W and Yiqing Xu (2024). 'LaLonde (1986) after Nearly Four Decades: Lessons Learned'. In: *arXiv preprint arXiv:2406.00827*.

Jacobs, Abigail Z and Hanna Wallach (2021). 'Measurement and fairness'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385.

Jaime, Sofia and Christoph Kern (2024). 'Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 237–253.

Jain, Shomik, Kathleen Creel and Ashia Camage Wilson (2024). 'Position: Scarce resource allocations that rely on machine learning should be randomized'. In: *Proceedings of the 41st International Conference on Machine Learning*, pp. 21148–21169.

Jaynes, Edwin T (1985). 'Some random observations'. In: *Synthese* 63, pp. 115–138.

Jeong, Haewon et al. (2022). 'Fairness without imputation: A decision tree approach for fair prediction with missing values'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9, pp. 9558–9566.

Jones, Camara Phyllis, Benedict I Truman, Laurie D Elam-Evans, Camille A Jones, Clara Y Jones, Ruth Jiles, Susan F Rumisha and Geraldine S Perry (2008). 'Using "socially assigned race" to probe white advantages in health status'. In: *Ethnicity & Disease* 18.4, pp. 496–504.

Jovanović, Nikola, Mislav Balunovic, Dimitar Iliev Dimitrov and Martin Vechev (2023). 'Fare: Provably fair representation learning with practical certificates'. In: *International Conference on Machine Learning*. PMLR, pp. 15401–15420.

Junquera, Álvaro F and Christoph Kern (2025). 'From rules to forests: Rule-based versus statistical models for jobseeker profiling'. In: *Journal for Labour Market Research* 59.1, pp. 1–27.

Kalluri, Pratyusha (2020). 'Don't ask if artificial intelligence is good or fair, ask how it shifts power'. In: *Nature* 583.7815, pp. 169–169.

Kang, Joseph DY and Joseph L Schafer (2007). 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data'. In: *Statistical Science* 22.4, pp. 523–539.

Karkkainen, Kimmo and Jungseock Joo (2021). 'Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558.

Kass, Robert E (2011). 'Statistical inference: The big picture'. In: *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 26.1, p. 1.

Kasy, Maximilian and Rediet Abebe (2021). 'Fairness, equality, and power in algorithmic decision-making'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576–586.

Kearns, Michael, Seth Neel, Aaron Roth and Zhiwei Steven Wu (2018). 'Preventing fairness gerrymandering: Auditing and learning for subgroup fairness'. In: *International Conference on Machine Learning*. PMLR, pp. 2564–2572.

Keiding, Niels and Thomas A Louis (2016). 'Perils and potentials of self-selected entry to epidemiological studies and surveys'. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.2, pp. 319–376.

Kelly, Markelle and Padhraic Smyth (2023). 'Variable-based calibration for machine learning classifiers'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 7, pp. 8211–8219.

Kern, Christoph, Ruben Bach, Hannah Mautner and Frauke Kreuter (2024). 'When small decisions have big impact: Fairness implications of algorithmic profiling schemes'. In: *ACM Journal on Responsible Computing* 1.4, pp. 1–30.

Kern, Christoph, Unai Fischer-Abaigar, Jonas Schweisthal, Dennis Frauen, Rayid Ghani, Stefan Feuerriegel, Mihaela van der Schaar and Frauke Kreuter (2025). 'Algorithms for reliable decision-making need causal reasoning'. In: *Nature Computational Science* 5.5, pp. 356–360.

Keswani, Vijay, Anay Mehrotra and L Elisa Celis (2024). 'Fair classification with partial feedback: An exploration-based data-collection approach'. In: *International Conference on Machine Learning*.

Keynes, John Maynard (1921). *A Treatise on Probability*. Macmillan & Co.

Khalili, Mohammad Mahdi, Xueru Zhang and Mahed Abroshan (2023). 'Loss balancing for fair supervised learning'. In: *International Conference on Machine Learning*. PMLR, pp. 16271–16290.

Khan, Khalil et al. (2015). 'Multi-class semantic segmentation of faces'. In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 827–831.

Khan, Zaid and Yun Fu (2021). 'One label, one billion faces: Usage and consistency of racial categories in computer vision'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 587–597.

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing and Bernhard Schölkopf (2017). 'Avoiding discrimination through causal reasoning'. In: *Advances in Neural Information Processing Systems* 30.

Kim, Kwangho and José R Zubizarreta (2023). 'Fair and robust estimation of heterogeneous treatment effects for policy learning'. In: *International Conference on Machine Learning*. PMLR, pp. 16997–17014.

Kim, Michael P, Amirata Ghorbani and James Zou (2019). 'Multiaccuracy: Black-box post-processing for fairness in classification'. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254.

Kitagawa, Toru and Aleksey Tetenov (2021). 'Equality-minded treatment choice'. In: *Journal of Business & Economic Statistics* 39.2, pp. 561–574.

Kleinberg, Bobby, Renato Paes Leme, Jon Schneider and Yifeng Teng (2023). 'U-calibration: Forecasting for an unknown agent'. In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, pp. 5143–5145.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2017). 'Inherent trade-offs in the fair determination of risk scores'. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pp. 43–1.

Kleinberg, Jon and Manish Raghavan (2021). 'Algorithmic monoculture and social welfare'. In: *Proceedings of the National Academy of Sciences* 118.22, e2018340118.

Knight, Frank Hyneman (1921). *Risk, Uncertainty and Profit*. Vol. 31. Houghton Mifflin.

Knittel, Marina, Max Springer, John P Dickerson and Mohammad Hajiaghayi (2023). 'Generalized reductions: Making any hierarchical clustering fair and balanced with low cost'. In: *International Conference on Machine Learning*. PMLR, pp. 17218–17242.

Kohavi, Ron (1996). 'Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid.' In: *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207.

Kolmogorov, Andrei Nikolaevich (1983). 'On logical foundations of probability theory'. In: *Probability Theory and Mathematical Statistics: Proceedings of the Fourth USSR-Japan Symposium, held at Tbilisi, USSR, August 23–29, 1982*. Ed. by K. Ito and J.V. Prokhorov. Springer, pp. 1–5.

Krippner, Greta R (2023). 'Unmasked: A history of the individualization of risk'. In: *Sociological Theory* 41.2, pp. 83–104.

Kroeber, Alfred L (1948). *Anthropology: Race, Language, Culture, Psychology, Prehistory*. Harcourt, Brace and Company.

Kuhn, Thomas S. (2012). 'Postscript'. In: *The Structure of Scientific Revolutions*. 4th edition. Chicago: University of Chicago Press.

Kumar, Aviral, Sunita Sarawagi and Ujjwal Jain (2018). 'Trainable calibration measures for neural networks from kernel mean embeddings'. In: *International Conference on Machine Learning*. PMLR, pp. 2805–2814.

Kuppler, Matthias, Christoph Kern, Ruben L Bach and Frauke Kreuter (2022). 'From fair predictions to just decisions? Conceptualizing algorithmic fair-

ness and distributive justice in the context of data-driven decision-making'. In: *Frontiers in Sociology* 7.

Kusuoka, Shigeo (2001). 'On law invariant coherent risk measures'. In: *Advances in Mathematical Economics*. Springer, pp. 83–95.

LaLonde, Robert J (1986). 'Evaluating the econometric evaluations of training programs with experimental data'. In: *The American economic review*, pp. 604–620.

Latour, Bruno (2015). 'Tarde's idea of quantification'. In: *The Social After Gabriel Tarde*. Routledge, pp. 187–202.

Le Guin, Ursula K (1969). *The Left Hand of Darkness*. Ace Books.

Lechner, Michael, Michael Knaus, Martin Huber, Markus Frölich, Stefanie Behncke, Giovanni Mellace and Anthony Strittmatter (2020). 'Swiss active labor market policy evaluation [Dataset]'. In: *Distributed by FORS*. DOI: https://doi.org/10.23662/FORS-DS-1203-1.

Lenhard, Johannes (2006). 'Models and statistical inference: The controversy between Fisher and Neyman–Pearson'. In: *The British Journal for the Philosophy of Ccience*.

Leonelli, Sabina (2019). 'Data governance is key to interpretation: Reconceptualizing data in data science'. In: *Harvard Data Science Review* 1.

— (2025). 'Environmental Intelligence: Redefining the philosophical premises of AI'. In: *Harvard Data Science Review* 7.4.

Lesko, Catherine R, Benjamin Ackerman, Michael Webster-Clark and Jessie K Edwards (2020). 'Target validity: Bringing treatment of external validity in line with internal validity'. In: *Current Epidemiology Reports* 7, pp. 117–124.

Levin, Daniel T (2000). 'Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit.' In: *Journal of Experimental Psychology: General* 129.4, p. 559.

Levin, Daniel T and Mahzarin R Banaji (2006). 'Distortions in the perceived lightness of faces: the role of race categories.' In: *Journal of Experimental Psychology: General* 135.4, p. 501.

Levy, Carl (2015). 'Racism, immigration and new identities in Italy'. In: *The Routledge Handbook of Contemporary Italy*. Routledge, pp. 49–63.

Lewis, David (1980). 'A subjectivist's guide to objective chance'. In: *Philosophical Papers (1986)* 2, 83—132.

— (1994). 'Humean supervenience debugged'. In: *Mind* 103.412, pp. 473–490.

Li, Danielle, Lindsey R Raymond and Peter Bergman (2020). *Hiring as exploration*. Tech. rep. National Bureau of Economic Research.

Li, Xinran and Xiao-Li Meng (2021). 'A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff'. In: *Journal of the American Statistical Association*.

Lipp, Robert (2005). 'Job seeker pProfiling'. In: *The Australian Experience*.

Little, Roderick JA (1986). 'Survey nonresponse adjustments for estimates of means'. In: *International Statistical Review/Revue Internationale de Statistique*, pp. 139–157.

Little, Roderick JA and Donald B Rubin (2020). *Statistical Analysis With Missing Data*. 3rd ed. John Wiley & Sons.

Liu, Lydia T, Inioluwa Deborah Raji, Angela Zhou, Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang, Hammaad Adam, Amanda Coston et al. (2025). 'Bridging prediction and intervention problems in social systems'. In: *arXiv preprint arXiv:2507.05216*.

Luccioni, Sasha, Christopher Akiki, Margaret Mitchell and Yacine Jernite (2024). 'Stable bias: Evaluating societal representations in diffusion models'. In: *Advances in Neural Information Processing Systems* 36.

Ludwig, David (2016). 'Ontological choices and the value-free ideal'. In: *Erkenntnis* 81.6, pp. 1253–1272.

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

Ma, Debbie S, Joshua Correll and Bernd Wittenbrink (2015). 'The Chicago face database: A free stimulus set of faces and norming data'. In: *Behavior Research Methods* 47, pp. 1122–1135.

Ma, Debbie S, Justin Kantner and Bernd Wittenbrink (2021). 'Chicago face database: Multiracial expansion'. In: *Behavior Research Methods* 53, pp. 1289–1300.

Macharia, Faith (2017). 'The Socio-cultural impacts of ethno-racism in Italy as functions of the global economic north/south divide'. In: *The Mellon Mays Undergraduate Fellowship Journal 2017*, p. 89.

Manski, Charles F (1990). 'Nonparametric bounds on treatment effects'. In: *The American Economic Review* 80.2, pp. 319–323.

— (1996). 'Learning about treatment effects from experiments with random assignment of treatments'. In: *Journal of Human Resources*, pp. 709–733.

— (2004). 'Statistical treatment rules for heterogeneous populations'. In: *Econometrica* 72.4, pp. 1221–1246.

Markus, Keith A (2021). 'Causal effects and counterfactual conditionals: Contrasting Rubin, Lewis and Pearl'. In: *Economics & Philosophy* 37.3, pp. 441–461.

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

McWatt, Tessa (2020). *Shame on Me: An Anatomy of Race and Belonging*. Random House Canada.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz et al. (2015). 'Identifying and cultivating superforecasters as a method of improving probabilistic predictions'. In: *Perspectives on Psychological Science* 10.3, pp. 267–281.

Memarrast, Omid, Linh Vu and Brian D Ziebart (2023). 'Superhuman fairness'. In: *International Conference on Machine Learning*. PMLR, pp. 24420–24435.

Meng, Xiao-Li (2018). 'Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election'. In: *The Annals of Applied Statistics* 12.2, pp. 685–726.

— (2022). 'Comments on "Statistical inference with non-probability survey samples"-Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples'. In: *Survey Methodology* 48.2, pp. 339–360.

Menon, Aditya Krishna and Robert C Williamson (2018). 'The cost of fairness in binary classification'. In: *Conference on Fairness, Accountability and Transparency*. PMLR, pp. 107–118.

Mercer, Andrew W, Frauke Kreuter, Scott Keeter and Elizabeth A Stuart (2017). 'Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference'. In: *Public Opinion Quarterly* 81.S1, pp. 250–271.

Merleau-Ponty, Maurice (1955). *Les aventures de la dialectique*. Gallimard.

Merler, Michele, Nalini Ratha, Rogerio S. Feris and John R. Smith (2019). 'Diversity in faces'. In: *arXiv preprint arXiv:1901.10436*.

Meta (2025). 'Personal superintelligence'. In: URL: https://www.meta.com/superintelligence/ (visited on 11/09/2025).

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh and Madeleine Clare Elish (2021). 'Algorithmic impact assessments and accountability: The co-construction of impacts'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735–746.

Metcalf, Jacob, Ranjit Singh, Emanuel Moss, Emnet Tafesse and Elizabeth Anne Watkins (2023). 'Taking algorithms to courts: A relational approach to algorithmic accountability'. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1450–1462.

Mickel, Jennifer (2024). 'Racial/ethnic categories in AI and algorithmic fairness: Why they matter and what they represent'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2484–2494.

MidJourney (2024). *MidJourney: AI-Powered Image Synthesis Platform*. Accessed: 2024-01-26. URL: https://www.midjourney.com.

Miettinen, Olli S (1985). *Theoretical Epidemiology*. Wiley.

Miles, Kierra L (2020). 'What Is Enough?: Understanding the Ostracization of Mixed People and How They Reaffirm Their Identity'. In: *The Macksey Journal* 1.1.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji and Timnit Gebru (2019). 'Model cards for model reporting'. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 220–229.

Monea, Alexander (2019). 'Race and computer vision'. In.

Moravec, Hans (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

Moss, Emanuel D (2021). 'The objective function: Science and society in the age of machine intelligence'. PhD thesis. City University of New York.

Mosselmans, Bert (2005). 'Adolphe Quetelet, the average man and the development of economic methodology'. In: *The European Journal of the History of Economic Thought* 12.4, pp. 565–582.

Munger, Kevin (2023). 'Temporal validity as meta-science'. In: *Research & Politics* 10.3.

Murphy, Allan H and Robert L Winkler (1977). 'Reliability of subjective probability forecasts of precipitation and temperature'. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 26.1, pp. 41–47.

Naeini, Mahdi Pakdaman, Gregory Cooper and Milos Hauskrecht (2015). 'Obtaining well calibrated probabilities using Bayesian binning'. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nelaturu, Sree Harsha, Nishaanth Kanna Ravichandran, Cuong Tran, Sara Hooker and Ferdinando Fioretto (2024). 'On The Fairness Impacts of Hardware Selection in Machine Learning'. In: *International Conference on Machine Learning*.

Neslen, Arthur (2021). 'Pushback against AI policing in Europe heats up over racism fears'. In: *Reuters*. URL: https://www.reuters.com/article/europe-tech-police-idINL8N2R92HQ/.

Neyman, Jerzy (1923/1990). 'On the application of probability theory to agricultural experiments. Essay on principles. Section 9.' In: *Statistical Science* 5.4. Trans. Dorota M. Dabrowska and Terence P. Speed, pp. 465–472.

Neyman, Jerzy and Karolina Iwaszkiewicz (1935). 'Statistical problems in agricultural experimentation'. In: *Supplement to the Journal of the Royal Statistical Society* 2.2, pp. 107–180.

Nielsen, Magnus Lindgaard, Jonas Skjold Raaschou-Pedersen, Emil Chrisander, David Dreyer Lassen, Julien Grenet, Anna Rogers and Andreas Bjerre-Nielsen (2025). 'Trading off performance and human oversight in algorithmic policy: Evidence from Danish college admissions'. In: *arXiv preprint arXiv:2411.15348*.

Nilipour, Leila (2021). 'Cynthia Arrieu-King's "The Betweens" offers refreshing, nuanced perspective on mixed-race identity'. In: *The Stanford Daily*. Accessed: 2025-01-02. URL: https://stanforddaily.com/2021/05/18/cynthia-arrieu-kings-the-betweens-offers-refreshing-nuanced-perspective-on-mixed-race-identity/.

Nixon, Jeremy, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel and Dustin Tran (2019). 'Measuring calibration in deep learning.' In: *CVPR Workshops*. Vol. 2. 7.

Noarov, Georgy and Aaron Roth (2024). *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/ (visited on 29/09/2025).

O'Neil, Cathy (2014). *Gillian Tett gets it very wrong on racial profiling*. Accessed: 2025-12-15. Mathbabe.org. URL: https://mathbabe.org/2014/08/25/gilian-tett-gets-it-very-wrong-on-racial-profiling/.

— (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Okati, Nastaran, Stratis Tsirtsis and Manuel Gomez Rodriguez (2023). 'On the within-group fairness of screening classifiers'. In: *International Conference on Machine Learning*. PMLR, pp. 26495–26516.

Okoro, Olihe N, Vibhuti Arya, Caroline A Gaither and Adati Tarfa (2021). 'Examining the inclusion of race and ethnicity in patient cases'. In: *American Journal of Pharmaceutical Education* 85.9, p. 8583.

Oosterloo, Serena and Gerwin van Schie (2018). 'The politics and biases of the "Crime Anticipation System" of the Dutch police'. In: *Proceedings of the In-*

ternational Workshop on Bias in Information, Algorithms, and Systems: Sheffield, United Kingdom, March 25, 2018. CEUR WS, pp. 30–41.

OpenAI (2023). *ChatGPT-4*. https://openai.com. Accessed: 2025-01-13.

OpenAI (2024). *DALL·E 3*. Accessed: 2025-01-26. URL: https://www.openai.com/dall-e-3.

Osanami Törngren, Sayaka (2020). 'Challenging the 'Swedish' and 'immigrant' dichotomy: How do multiracial and multi-ethnic Swedes identify themselves?' In: *Journal of Intercultural Studies* 41.4, pp. 457–473.

Osanami Törngren, Sayaka, Nahikari Irastorza and Dan Rodríguez-García (2021). 'Understanding multiethnic and multiracial experiences globally: Towards a conceptual framework of mixedness'. In: *Journal of Ethnic and Migration Studies* 47.4, pp. 763–781.

Oxford English Dictionary (2025a). *Mixed-race*. Accessed: 28-Jan-2025. URL: https://www.oed.com/search/dictionary/?scope=Entries&q=mixed-race&tl=true.

— (2025b). *mulatto, n. & adj.* Accessed: January 12, 2025. URL: https://www.oed.com/dictionary/mulatto_n.

Ozturk, Kagan, Haiyu Wu and Kevin W Bowyer (2024). 'Can the accuracy bias by facial hairstyle be reduced through balancing the training data?' In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1519–1528.

Park, Sungho and Hyeran Byun (2024). 'Fair-VPT: Fair visual prompt tuning for image classification'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12268–12278.

Pearl, Judea (2009). *Causality*. Cambridge University Press.

Pearl, Judea and Elias Bareinboim (2014). 'External validity: From do-calculus to transportability across populations'. In: *Statistical Science* 29.4, pp. 579–595.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pedreshi, Dino, Salvatore Ruggieri and Franco Turini (2008). 'Discrimination-aware data mining'. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568.

Peirce, Charles Sanders (1878). 'The doctrine of chances'. In: 12, pp. 604–615.

Peltonen, Jaakko, Wen Xu, Timo Nummenmaa and Jyrki Nummenmaa (2023). 'Fair neighbor embedding'. In: *International Conference on Machine Learning*. PMLR, pp. 27564–27584.

Perdomo, Juan Carlos (2024). 'The relative value of prediction in algorithmic decision making'. In: *International Conference on Machine Learning*, pp. 40439–40460.

— (2025). 'Revisiting the predictability of performative, social events'. In: *International Conference on Machine Learning*.

Perdomo, Juan Carlos, Tolani Britton, Moritz Hardt and Rediet Abebe (2025). 'Difficult lessons on social prediction from Wisconsin public schools'. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2682–2704.

Perdomo, Juan Carlos and Benjamin Recht (2025). 'In defense of defensive forecasting'. In: *arXiv preprint arXiv:2506.11848*.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner and Moritz Hardt (2020). 'Performative prediction'. In: *International Conference on Machine Learning*. PMLR, pp. 7599–7609.

Phillips, P Jonathon, Hyeonjoon Moon, Syed A Rizvi and Patrick J Rauss (2000). 'The FERET evaluation methodology for face-recognition algorithms'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.10, pp. 1090–1104.

Poisson, Siméon-Denis (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier.

Polanyi, Michael (1958). *Personal Knowledge*. University of Chicago Press.

Poole, Deborah (2005). 'An excess of description: Ethnography, race, and visual technologies'. In: *Annual Review of Anthropology* 34.1, pp. 159–179.

Popper, Karl R (1959). 'The propensity interpretation of probability'. In: *The British Journal for the Philosophy of Science* 10.37, pp. 25–42.

Porter, Theodore M (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton University Press.

— (1995). *Trust in Numbers*. Princeton University Press.

Potochnik, Angela (2017). *Idealization and the Aims of Science*. University of Chicago Press.

ProPublica (2016). *Compas analysis github repository*. URL: https://github.com/propublica/compas-analysis.

Qraitem, Maan, Kate Saenko and Bryan A Plummer (2023). 'Bias mimicking: A simple sampling approach for bias mitigation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20320.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Saurabh Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. (2021). 'Learning transferable visual models from natural language supervision'. In: *International Conference on Machine Learning*.

Raghavan, Manish and Pauline T Kim (2024). 'Limitations of the "four-fifths rule" and statistical parity tests for measuring fairness'. In: *Georgetown Law Technology Review* 8, p. 93.

Ramsey, Frank P (1926/1931). 'Truth and probability'. In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. Routledge. Chap. VII, pp. 156–198.

— (1928/1931). 'Further considerations'. In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. Routledge. Chap. VIII, pp. 199–211.

Ravishankar, Pavan, Qingyu Mo, Edward McFowland III and Daniel B Neill (2023). 'Provable detection of propagating sampling bias in prediction models'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 8, pp. 9562–9569.

Recht, Ben (2025a). *There Is No Data-Generating Distribution*. Accessed: 2025-12-20. URL: https://www.argmin.net/p/there-is-no-data-generating-distribution.

Recht, Benjamin (2025b). 'The actuary's final word on algorithmic decision making'. In: *arXiv preprint arXiv:2509.04546*.

Redmond, Michael and Alok Baveja (2002). 'A data-driven software tool for enabling cooperative information sharing among police departments'. In: *European Journal of Operational Research* 141.3, pp. 660–678.

Reichenbach, Hans (1949). *The Theory of Probability*. University of California Press.

Remedios, Jessica D and Alison L Chasteen (2013). 'Finally, someone who "gets" me! Multiracial people value others' accuracy about their race.' In: *Cultural Diversity and Ethnic Minority Psychology* 19.4, p. 453.

Ricanek, Karl and Tamirat Tesafaye (2006). 'Morph: A longitudinal image database of normal adult age-progression'. In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, pp. 341–345.

Rich, Camille Gear (2013). 'Elective race: Recognizing race discrimination in the era of racial self-identification'. In: *Geo. LJ* 102, p. 1501.

Rietz, HL (1924). 'On certain topics in the mathematical theory of statistics'. In: *Bulletin of the American Mathematical Society* 30.8, pp. 417–453.

— (1932). 'On the Lexis theory and the analysis of variance'. In: *Bulletin of the American Mathematical Society* 38.10, pp. 731–735.

Robbin, Alice (2000). 'Classifying racial and ethnic group data in the United States: The politics of negotiation and accommodation'. In: *Journal of Government Information* 27.2, pp. 129–156.

Robins, James M (1989). 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies'. In: *Health Service Research Methodology: A Focus On AIDS*, pp. 113–159.

Robins, James M and Andrea Rotnitzky (1995). 'Semiparametric efficiency in multivariate regression models with missing data'. In: *Journal of the American Statistical Association* 90.429, pp. 122–129.

Rockafellar, R Tyrrell and Stan Uryasev (2013). 'The fundamental risk quadrangle in risk management, optimization and statistical estimation'. In: *Surveys in Operations Research and Management Science* 18.1-2, pp. 33–53.

Rodríguez-García, Dan, Miguel Solana, Anna Ortiz and Beatriz Ballestín (2021). 'Blurring of colour lines? Ethnoracially mixed youth in Spain navigating identity'. In: *Journal of Ethnic and Migration Studies* 47.4, pp. 838–860.

Roh, Yuji, Kangwook Lee, Steven Euijong Whang and Changho Suh (2023). 'Improving fair training under correlation shifts'. In: *International Conference on Machine Learning*. PMLR, pp. 29179–29209.

Rose, Evan K (2023). 'A constructivist perspective on empirical discrimination research'. In: *Journal of Economic Literature* 61.3, pp. 906–923.

Rosenbaum, Paul R and Donald B Rubin (1983). 'The central role of the propensity score in observational studies for causal effects'. In: *Biometrika* 70.1, pp. 41–55.

Ross, Casey (2022). 'Epic's overhaul of a flawed algorithm shows Why AI oversight is a life-or-Death issue'. en. In: *STAT*. Accessed: 2025-12-11. URL: https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/.

Roth, Aaron and Alexander Tolbert (2025). 'Resolving the reference class problem at scale'. In: *Philosophy of Science* 92.4, pp. 868–882.

Roth, Aaron, Alexander Tolbert and Scott Weinstein (2023). 'Reconciling individual probability forecasts'. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 101–110.

Roth, Wendy D (2016). 'The multiple dimensions of race'. In: *Ethnic and Racial Studies* 39.8, pp. 1310–1338.

Rubin, Donald B (1974). 'Estimating causal effects of treatments in randomized and nonrandomized studies'. In: *Journal of Educational Psychology* 66.5, p. 688.

— (1976). 'Inference and missing data'. In: *Biometrika* 63.3, pp. 581–592.

— (1990). 'Comment: Neyman (1923) and causal inference in experiments and observational studies'. In: *Statistical Science* 5.4, pp. 472–480.

Rudin, Cynthia, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen and Zachery Boner (2024). 'Position: Amazing things come from having many good models'. In: *International Conference on Machine Learning*.

Rudner, Richard (1953). 'The scientist qua scientist makes value judgments'. In: *Philosophy of Science* 20.1, pp. 1–6.

Rudolph, Jacqueline E, Yongqi Zhong, Priya Duggal, Shruti H Mehta and Bryan Lau (2023). 'Defining representativeness of study samples in medical and population health research'. In: *BMJ Medicine* 2.1.

Sachdeva, Silonie (2009). 'Fitzpatrick skin typing: Applications in dermatology'. In: *Indian Journal of Dermatology, Venereology and Leprology* 75, p. 93.

Salmon, Wesley C (1966). *The Foundations of Scientific Inference*. University of Pittsburgh Press.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh and Lora M Aroyo (2021). '"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI'. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Sanders, Frederick (1963). 'On subjective probability forecasting'. In: *Journal of Applied Meteorology and Climatology* 2.2, pp. 191–201.

Savage, Leonard J (1972). *The Foundations of Statistics*. 2nd edition. John Wiley and Sons.

Schervish, Mark J (1985). 'Discussion: Calibration-based empirical probability'. In: *The Annals of Statistics* 13.4, pp. 1274–1282.

Schnorr, Claus P (2007). *Zufälligkeit und Wahrscheinlichkeit: Eine Algorithmische Begründung der Wahrscheinlichkeitstheorie*. Springer.

Searle, John R (2006). 'Social ontology: Some basic principles'. In: *Anthropological Theory* 6.1, pp. 12–29.

Seidenfeld, Teddy, Mark J Schervish and Joseph B Kadane (2012). 'Forecasting with imprecise probabilities'. In: *International Journal of Approximate Reasoning* 53.8, pp. 1248–1261.

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian and Janet Vertesi (2019). 'Fairness and abstraction in sociotechnical systems'. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 59–68.

Semenova, Lesia, Harry Chen, Ronald Parr and Cynthia Rudin (2023). 'A path to simpler models starts with noise'. In: *Advances in Neural Information Processing Systems* 36.

Sen, Amartya (2010). *The Idea of Justice*. Penguin books.

Shafer, Glenn and Vladimir Vovk (2008). 'A tutorial on conformal prediction'. In: *Journal of Machine Learning Research* 9.3.

— (2019). *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons.

Sharma, Vibhhu and Bryan Wilder (2025). 'Comparing targeting strategies for maximizing social welfare with limited resources'. In: *The Thirteenth International Conference on Learning Representations*.

Shirali, Ali, Rediet Abebe and Moritz Hardt (2024). 'Allocation requires prediction only if inequality is low'. In: *International Conference on Machine Learning*, pp. 45114–45153.

Simon, Patrick (2015). 'The choice of ignorance: The debate on ethnic and racial statistics in France'. In: *Social Statistics and Ethnic Diversity: Cross-national Perspectives in Classifications and Identity Politics*, pp. 65–87.

— (2017). 'The failure of the importation of ethno-racial statistics in Europe: Debates and controversies'. In: *Ethnic and Racial Studies* 40.13, pp. 2326–2332.

Singh, Harvineet, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara and Chris Russell (2023). 'When do minimax-fair learning and empirical risk minimization coincide?' In: *International Conference on Machine Learning*. PMLR, pp. 31969–31989.

Sloane, Mona, Emanuel Moss, Olaitan Awomolo and Laura Forlano (2022). 'Participation is not a design fix for machine learning'. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–6.

Smedley, Audrey and Brian D Smedley (2005). 'Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race.' In: *American Psychologist* 60.1, p. 16.

Soen, Alexander, Hisham Husain and Richard Nock (2023). In: *International Conference on Machine Learning*. PMLR, pp. 32105–32144.

Speicher, Till, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller and Muhammad Bilal Zafar (2018). 'A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices'. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248.

Spiegelhalter, David (2024). 'Does probability exist?' In: *Nature* 636, p. 19.

Spirtes, Peter, Clark N Glymour and Richard Scheines (2000). *Causation, Prediction, and Search*. MIT Press.

Stanley, Jay (2018). 'New Orleans program offers lessons ln pitfalls Of predictive policing'. In: *American Civil Liberties Union*. Accessed: 2025-12-15. URL:

https://www.aclu.org/news/privacy-technology/new-orleans-program-offers-lessons-pitfalls-predictive-policing.

Steckler, Allan and Kenneth R McLeroy (2008). 'The importance of external validity'. In: *American Journal of Public Health* 98.1, pp. 9–10.

Stein, Michael Isaac (2020). 'New Orleans city council bans facial recognition, predictive policing and other surveillance tech'. In: *The Lens*. Accessed: 2025-12-15. URL: https://thelensnola.org/2020/12/18/new-orleans-city-council-approves-ban-on-facial-recognition-predictive-policing-and-other-surveillance-tech/.

Stock, Michelle L, Frederick X Gibbons, Janine B Beekman, Kipling D Williams, Laura S Richman and Meg Gerrard (2018). 'Racial (vs. self) affirmation as a protective mechanism against the effects of racial exclusion on negative affect and substance use vulnerability among black young adults'. In: *Journal of Behavioral Medicine* 41, pp. 195–207.

Streater, RF (2000). 'Classical and quantum probability'. In: *Journal of Mathematical Physics* 41.6, pp. 3556–3603.

Strevens, Michael (2006). 'Probability and chance'. In: *Encyclopedia of Philosophy, second edition. Macmillan Reference USA, Detroit*.

Strikwerda, Litska (2021). 'Predictive policing: The risks associated with risk assessment'. In: *The Police Journal* 94.3, pp. 422–436.

Tett, Gillian (2014). 'Mapping crime–or stirring hate?' In: *Financial Times*. URL: https://www.ft.com/content/200bebee-28b9-11e4-8bda-00144feabdc0.

Thoma, Johanna (2019). 'Risk aversion and the long run'. In: *Ethics* 129.2, pp. 230–253.

Thompson, Debra (2013). 'Making (mixed-)race: Census politics and the emergence of multiracial multiculturalism in the United States, Great Britain and Canada'. In: *Accounting for Ethnic and Racial Diversity*. Routledge, pp. 53–70.

Thorp, Edward O (1998). 'The invention of the first wearable computer'. In: *Digest of Papers. Second International Symposium on Wearable Computers*. IEEE, pp. 4–8.

Tifrea, Alexandru, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami and Flavien Prost (2024). 'FRAPPÉ: A Group Fairness Framework for Post-Processing Everything'. In: *International Conference on Machine Learning*.

Travers, Eoin, Merle T Fairhurst and Ophelia Deroy (2020). 'Racial bias in face perception is sensitive to instructions but not introspection'. In: *Consciousness and Cognition* 83, p. 102952.

Trehan, Nidhi and Angéla Kóczé (2009). 'Racism, (neo-) colonialism, and social justice: The struggle for the soul of the Romani movement in post-socialist Europe'. In: *Racism Postcolonialism Europe*, pp. 50–73.

US Office of Management and Budget (1997). *OMB Statistical Policy Directive No. 15: 1997 Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity*. URL: https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf.

Union, European (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council*. Accessed: 2025-01-09. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679.

United States Census Bureau (n.d.). *PUMS Data*. https://www.census.gov/programs-surveys/acs/microdata/access.html. Accessed: 2025-01-21.

— (2021). *Improved Race and Ethnicity Measures Reveal United States Population Is Much More Multiracial*. https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html. Accessed: 2025-01-11.

Vafa, Keyon (2024). 'Is causal inference compatible with frictionless reproducibility?' In: *Harvard Data Science Review* 6.1.

Vaicenavicius, Juozas, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll and Thomas Schön (2019). 'Evaluating model calibration in classification'. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3459–3467.

Vaihinger, Hans (1911). *Die Philosophie des Als Ob. System der theoretischen, praktischen und religiösen Fiktionen der Menschheit auf Grund eines idealistischen Positivismus. Mit einem Anhang über Kant und Nietzsche*. Reuther & Reichard.

Van Fraassen, Bas C (1983). 'Calibration: A frequency justification for personal probability'. In: *Physics, Philosophy and Psychoanalysis: Essays in Honour of Adolf Grünbaum*, pp. 295–319.

Vapnik, Vladimir (1982). *Estimation of Dependences Based on Empirical Data*. Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). 'Attention is all you need'. In: *Advances in Neural Information Processing Systems* 30.

Vela, Daniel, Andrew Sharp, Richard Zhang, Trang Nguyen, An Hoang and Oleg S Pianykh (2022). 'Temporal quality degradation in AI models'. In: *Scientific Reports* 12.1, p. 11654.

Venn, John (1866). *The Logic of Chance*. Macmillan.

Vovk, Vladimir (2006). 'Predictions as statements and decisions'. In: *arXiv preprint cs/0606093*.

Vovk, Vladimir and Glenn Shafer (2025). 'A conversation with A. Philip Dawid'. In: *Statistical Science* 40.1, pp. 148–166.

Vovk, Vladimir, Akimichi Takemura and Glenn Shafer (2005). 'Defensive forecasting'. In: *AISTATS*. Vol. 2005, pp. 365–372.

Vredenburgh, Kate (2022). 'Fairness'. In: *The Oxford Handbook of AI Governance*. Ed. by Justin B Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M Hudson, Anton Korinek, Matthew M Young and Baobao Zhang.

Wachter, Sandra, Brent Mittelstadt and Chris Russell (2021). 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI'. In: *Computer Law & Security Review* 41, p. 105567.

Wakker, Peter (1994). 'Separating marginal utility and probabilistic risk aversion'. In: *Theory and Decision* 36.1, pp. 1–44.

Walley, Peter (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman-Hall.

Wang, Angelina, Sayash Kapoor, Solon Barocas and Arvind Narayanan (2024). 'Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy'. In: *ACM Journal on Responsible Computing* 1.1, pp. 1–45.

Wang, Mei and Weihong Deng (2020). 'Mitigating bias in face recognition using skewness-aware reinforcement learning'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9322–9331.

Wang, Mei, Weihong Deng, Jiani Hu, Xunqiang Tao and Yaohai Huang (2019). 'Racial faces in the wild: Reducing racial bias by information maximization adaptation network'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 692–702.

Wasserman, Larry (2003). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Watson-Daniels, Jamelle, David C Parkes and Berk Ustun (2023). 'Predictive multiplicity in probabilistic classification'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.9, pp. 10306–10314.

Westen, Peter (1982). 'The empty idea of equality'. In: *Harvard Law Review* 95.3, pp. 537–596.

Westin, Charles (2003). 'Young people of migrant origin in Sweden'. In: *International Migration Review* 37.4, pp. 987–1010.

Westreich, Daniel, Jessie K Edwards, Catherine R Lesko, Stephen R Cole and Elizabeth A Stuart (2019). 'Target validity and the hierarchy of study designs'. In: *American Journal of Epidemiology* 188.2, pp. 438–443.

Wieringa, Maranke (2020). 'What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18.

Wightman, Linda F (1998). 'LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series'. In: *LSAC Research Report Series*. Accessed: 2025-01-15. URL: https://archive.lawschooltransparency.com/reform/projects/investigations/2015/documents/NLBPS.pdf.

Wilkens, Andreas (2021). *Vorausschauende Polizeiarbeit: Bayerns Polizei beendet die Arbeit mit PRECOBS*. Accessed: 2025-12-28. heise online. URL: https://www.heise.de/news/Vorausschauende-Polizeiarbeit-Bayerns-Polizei-beendet-die-Arbeit-mit-PRECOBS-6233501.html.

Williams, Donald (1947). *The Ground of Induction*. Harvard University Press.

Williams, Kim M (2003). 'From civil rights to the multiracial movement'. In: *New faces in a changing America: Multiracial identity in the 21st century*, pp. 85–98.

Williamson, Robert and Aditya Menon (2019). 'Fairness risk measures'. In: *International Conference on Machine Learning*. PMLR, pp. 6786–6797.

Wilson, James F, Michael E Weale, Alice C Smith, Fiona Gratrix, Benjamin Fletcher, Mark G Thomas, Neil Bradman and David B Goldstein (2001). 'Population genetic structure of variable drug response'. In: *Nature Genetics* 29.3, pp. 265–269.

Wu, Changbao (2022). 'Statistical inference with non-probability survey samples'. In: *Survey Methodology* 48.2, pp. 283–311.

Xian, Ruicheng, Qiaobo Li, Gautam Kamath and Han Zhao (2024). 'Differentially private post-processing for fair regression'. In: *International Conference on Machine Learning*, pp. 54212–54235.

Xian, Ruicheng, Lang Yin and Han Zhao (2023). 'Fair and optimal classification via post-processing'. In: *International Conference on Machine Learning*. PMLR, pp. 37977–38012.

Xu, Gezheng, Qi Chen, Charles X Ling, Boyu Wang and Changjian Shui (2024). 'Intersectional unfairness discovery'. In: *International Conference on Machine Learning*, pp. 54888–54917.

Xu, Shizhou and Thomas Strohmer (2023). 'Fair data representation for machine learning at the Pareto frontier'. In: *Journal of Machine Learning Research* 24.331, pp. 1–63.

York, Richard and Brett Clark (2007). 'The problem with prediction: Contingency, emergence, and the reification of projections'. In: *The Sociological Quarterly* 48.4, pp. 713–743.

Yucer, Seyma, Amir Atapour Abarghouei, Noura Al Moubayed and Toby P Breckon (2024a). 'Disentangling racial phenotypes: Fine-grained control of race-related facial phenotype characteristics'. In: *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–10.

Yucer, Seyma, Samet Akçay, Noura Al-Moubayed and Toby P Breckon (2020). 'Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–19.

Yucer, Seyma, Furkan Tektas, Noura Al Moubayed and Toby Breckon (2024b). 'Racial bias within face recognition: A survey'. In: *ACM Computing Surveys* 57.4, pp. 1–39.

Yucer, Seyma, Furkan Tektas, Noura Al Moubayed and Toby P. Breckon (2022). 'Measuring hidden bias within face recognition via racial phenotypes'. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 3202–3211. DOI: 10.1109/wacv51458.2022.00326. URL: http://dx.doi.org/10.1109/WACV51458.2022.00326.

Yudell, Michael, Dorothy Roberts, Rob DeSalle and Sarah Tishkoff (2016). 'Taking race out of human genetics'. In: *Science* 351.6273, pp. 564–565.

Zadrozny, Bianca and Charles Elkan (2001). 'Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers'. In: *International Conference on Machine Learning*. PMLR, pp. 609–616.

Zezulka, Sebastian and Konstantin Genin (2024). 'From the fair distribution of predictions to the fair distribution of social goods: Evaluating the impact of fair machine learning on long-term unemployment'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1984–2006.

Zhang, Bowen, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang and Baining Guo (2022). 'Styleswin: Transformer-based GAN for high-resolution image generation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11304–11314.

Zhang, Lujing, Aaron Roth and Linjun Zhang (2024). 'Fair risk control: A generalized framework for calibrating multi-group fairness risks'. In: *International Conference on Machine Learning*.

Zhang, Zhifei, Yang Song and Hairong Qi (2017). 'Age progression/regression by conditional adversarial autoencoder'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5810–5818.

Zhang, Zhiwei, Zhen Chen, James F Troendle and Jun Zhang (2012). 'Causal inference on quantiles with an obstetric application'. In: *Biometrics* 68.3, pp. 697–706.

Zhao, Dora, Jerone TA Andrews, Orestis Papakyriakopoulos and Alice Xiang (2024). 'Position: Measure dataset diversity, don't just claim it'. In: *Forty-first International Conference on Machine Learning*.

Zhao, Shengjia, Michael Kim, Roshni Sahoo, Tengyu Ma and Stefano Ermon (2021). 'Calibrating predictions to decisions: A novel approach to multi-class calibration'. In: *Advances in Neural Information Processing Systems* 34, pp. 22313–22324.

Zhou, Mi, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim and Kannan Srinivasan (2024). 'Bias in generative AI'. In: *arXiv preprint arXiv:2403.02726*.

Zhu, Zhaowei, Yuanshun Yao, Jiankai Sun, Hang Li and Yang Liu (2023). 'Weak proxies are sufficient and preferable for fairness with missing sensitive attributes'. In: *International Conference on Machine Learning*. PMLR, pp. 43258–43288.

Ziegler, Daniel M, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano and Geoffrey Irving (2020). 'Fine-tuning language models from human preferences'. In: *arXiv preprint arXiv:1909.08593*.