Forthcoming in *Analysis*

Diving into Fair Pools: Algorithmic Fairness, Ensemble Forecasting, and the Wisdom of Crowds

Rush Stewart[1] and Lee Elkin[2]

[1]University of Rochester
[2]The Alan Turing Institute

February 18, 2025

Abstract

Is the pool of fair predictive algorithms fair? It depends, naturally, on both the criteria of fairness and on how we pool. We catalog the relevant facts for some of the most prominent statistical criteria of algorithmic fairness and the dominant approaches to pooling forecasts: linear, geometric, and multiplicative. Only linear pooling, a format at the heart of ensemble methods, preserves any of the central criteria we consider. Drawing on work in the social sciences and social epistemology on the theoretical foundations of the wisdom of crowds, we explain how our observations present an exception to the general trend of finding tradeoffs between the accuracy and fairness of forecasts.

Keywords. Accuracy-fairness trade-off; AI ethics; algorithmic fairness; Crowds Beat Averages Law; opinion pooling

1 Introduction

Algorithms play a number of important roles in our private and public lives. They generate search engine results, organize news feeds on social media, and identify promising romantic partners. They inform judicial, loan, social benefits, and college admissions decisions. They also raise pressing and vexing ethical challenges. For instance, some algorithms used in the US criminal justice system predict whether individuals will recidivate. Famously, such algorithms have been found to exhibit apparent race- and sex-based bias like rating black non-recidivists as significantly more like to re-offend than white non-recidivists (Angwin et al., 2016a,b). Partially in reaction to such findings, the study of algorithmic fairness has assumed a prominent role in computer science, philosophy, and other fields.

A theoretically interesting and ethically salient finding that has emerged from these studies is that implementing sensible notions of *fairness* can come at the cost of *accuracy* (Corbett-Davies et al., 2017; Menon and Williamson, 2018; Kearns and Roth, 2019).

Grgić-Hlača et al. (2017) consider this finding in the setting of ensemble methods. In ensemble classification, a set of classifiers, rather than a single classifier, is produced. These classifiers are then combined to yield a classifier that is, in many cases, superior in some sense (Leutbecher and Palmer, 2008; Rokach, 2010). The goal of ensemble methods is often to improve accuracy. Types of ensemble methods for classifiers that are popular in machine learning and artificial intelligence research include bagging and boosting (Maclin and Opitz, 1997; Bühlmann, 2012). Bagging, for instance, generates a collection of individual or base classifiers by training some model-fitting method on different sequences of random samples from the training data. These classifiers can then be aggregated to form an ensemble classifier with reduced variance and mean squared error (Bühlmann, 2012). Among the findings that Grgić-Hlača and coauthors report are that ensemble classifiers preserve some notions of fairness and that they "can achieve better accuracy-fairness trade-offs than a single classifier" (2017: 1).

While the focus of Grgić-Hlača and coauthors is on binary classifiers, ours will be on more general probabilistic forecasts. A large number of algorithmic fairness criteria have been proposed for this setting. We consider ten such criteria. In Section 3, we observe, first, that an ensemble formed by linear pooling preserves several criteria of central or recent concern in work on algorithmic fairness and fails to preserve others. Second, unlike Grgić-Hlača and coauthors, we consider geometric and multiplicative pooling, prominent alternative ways of aggregating probabilities. Both geometric and multiplicative pooling fail to preserve any of the ten criteria we discuss. This catalog is one of the main contributions of our essay, and contributes data to debates regarding the appeal of various fairness criteria or approaches to pooling forecasts. Third, appealing to the Crowds Beat Averages Law, we observe in Section 4 that the inaccuracy of the linear pool is no greater than the average inaccuracy of the individual assessors. While the accuracy benefits of ensemble methods are widely-appreciated in computer science and machine learning, our route to this point emphasizes a connection to the social science and social epistemology literature on the wisdom of crowds. At least for several metrics of fairness and linear pooling, ensemble methods present a sort of exception to the trend of finding tradeoffs between the two aims of accuracy and fairness for predictive algorithms (Kearns and Roth, 2019). We comment on the dialectical significance of our observations in Section 5.

2 Algorithmic Fairness

Predictive algorithms can be thought of as making forecasts about whether individuals in some relevant population have a particular property y of concern. The property might be whether or not the individual will recidivate or whether or not she or he will default on a loan, for example. Let N be a finite population of individuals. Associated with each individual $i \in N$ typically is a vector of i's features or characteristics. In general the relevant features depend on the prediction task. It might include type of crime for predicting recidivism or annual income in predicting loan default. Since these vectors will not play a central role in our discussion, we will abstract from them, not representing them formally. Let $Y : N \to \{0, 1\}$ be a random variable representing whether an individual i has property y: $Y(i) = 1$ if i has y, and $Y(i) = 0$ otherwise. The population N can be divided into groups

in various ways. Concern with fair treatment is typically concern for fair treatment across salient, sensitive social groupings: race, sex, disability status, and so on. Let π be a partition of N representing some social grouping. Each individual in N belongs to exactly one cell of π.

An *assessor* $h : N \to [0, 1]$ assigns each individual i a number in the unit interval which can be interpreted as the probability that $Y(i) = 1$, that i has property y. In order to talk about proportions in groups and to formulate certain central fairness criteria, we also introduce a uniform probability distribution P on N. For any group $G \subseteq N$, we have $P(G) = \sum_{i \in G} P(i)$, and $P_G = P(\cdot|G)$ is uniform on G. The *base rate* (or prevalence) of y in G is $\mu_G = P_G(Y = 1)$. In words, the base rate in G is the proportion of individuals in G that have property y. The distribution P also allows us to define expectations. For example, $E_G(h) = E(h|G)$ is the expected value of h in group G; since P_G is uniform, this is just the average value of h in G.

The interesting ethical question here concerns the properties that h should have in order to be considered fair. The standard approach investigates potential necessary criteria of fairness. A rapidly expanding set of statistical criteria have been introduced and studied in the literature (Narayanan, 2018; Verma and Rubin, 2018; Eva, 2022; Nielsen and Stewart, 2024; Stewart, 2024). A sizeable subset of this expanding literature on criteria of algorithmic fairness reports results establishing that many sets of these criteria are effectively impossible to jointly satisfy (Borsboom et al., 2008; Chouldechova, 2017; Kleinberg et al., 2017; Stewart et al., 2024). The influential result of Kleinberg and coauthors, for instance, establishes that two compelling criteria—Calibration and Equalized Odds—are inconsistent except in unrealistic cases in which the base rates for all groups are identical or the assessor is perfect (Kleinberg et al., 2017). This observation has motivated a search for ways of modifying Calibration and Equalized Odds. Here, we will consider ten of these statistical criteria of fairness that are of core or recent interest. They can be divide into three groups: criteria related to Calibration, criteria related to Equalized Odds, and criteria related to a property called Statistical Parity. This division mirrors the one in (Barocas et al., 2023: 61). Both because of space limitations and because it has been done a number of times elsewhere in the literature (e.g., Barocas et al., 2023), we refrain from rehearsing explanations and motivations for these properties.

First, we have Calibration—the "dominant fairness criterion" in the literature according to some (Corbett-Davies et al., 2017: 799)—and various logically independent ways of relaxing it (Stewart et al., 2024; Nielsen and Stewart, 2024).

> **Calibration**. For an assessor h of N and any group G in the relevant partition π of N, $P_G(Y = 1|h = p) = p$ for all $p \in [0, 1]$ such that $P_G(h = p) > 0$.
>
> **Base Rate Tracking**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $E_G(h) - \mu_G = E_{G'}(h) - \mu_{G'}$.
>
> **Ratio Base Rate Tracking**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $\frac{E_G(h)}{\mu_G} = \frac{E_{G'}(h)}{\mu_{G'}}$ (whenever both ratios are defined).
>
> **Spanning**. For an assessor h of N and any group G in the relevant partition π of N, the base rate μ_G lies in the interval $[\min_{i \in G} h(i), \max_{i \in G} h(i)]$.

Predictive Equity. For an assessor h of N and any groups G, G' in a partition π of N, $P_G(Y = 1|h = p) = P_{G'}(Y = 1|h = p)$ for all $p \in [0, 1]$ such that $P_G(h = p), P_{G'}(h = p) > 0$.

Next, we have Equalized Odds and two other criteria that often go by that name in the literature (Grant, 2023). We follow the naming conventions adopted in (Nielsen and Stewart, 2024). Strong Equalized Odds implies both Equalized Odds and Threshold Equalized Odds, but those two properties are logically independent of each other.

Strong Equalized Odds. For an assessor h of N and any groups G, G' in the relevant partition π of N, $P_G(h = p|Y = 0) = P_{G'}(h = p|Y = 0)$ and $P_G(h = p|Y = 1) = P_{G'}(h = p|Y = 1)$ for any p (whenever those probabilities are defined).

Equalized Odds. For an assessor h of N and any groups G, G' in the relevant partition π of N, $E_G(h|Y = 0) = E_{G'}(h|Y = 0)$ and $E_G(h|Y = 1) = E_{G'}(h|Y = 1)$ (whenever those terms are defined).

Threshold Equalized Odds. For an assessor h of N, a specified threshold $t \in [0, 1]$, and any groups G, G' in the relevant partition π of N, $P_G(h > t|Y = 0) = P_{G'}(h > t|Y = 0)$ and $P_G(h < t|Y = 1) = P_{G'}(h < t|Y = 1)$ for any p (whenever those probabilities are defined).

Finally, we will also consider the following parity properties (Räz, 2021).

Strong Statistical Parity. For an assessor h of N and any groups G, G' in the relevant partition π of N, $P_G(h = p) = P_{G'}(h = p)$ for any $p \in [0, 1]$.

Statistical Parity. For an assessor h of N and any groups G, G' in the relevant partition π of N, $E_G(h) = E_{G'}(h)$.

It is not difficult to show that Strong Statistical Parity implies Statistical Parity, which we take to justify our naming convention. While Statistical Parity is similar to Equalized Odds, formulated basically without conditioning on the positive or negative class, the two properties are logically independent.

3 Pooling and Preserving Fairness

Various ways of aggregating a set of probabilistic forecasts have been studied in depth in economics (Mongin, 1995), philosophy (Dietrich and List, 2016; Elkin and Pettigrew, 2025), statistics (Genest and Zidek, 1986), and other fields of research. Pooling methods are ensemble methods for probability assessments. Among the most prominent and best supported means of aggregating probabilities is the linear pool. A number of considerations in favor of linear pooling have been adduced, considerations having to do with informational parsimony (McConway, 1981), accuracy (Pettigrew, 2019), and Pareto and coherence conditions (Gilboa et al., 2004; Nielsen, 2019). Linear averages are at the core of ensemble methods, as

Bühlmann notes: "the general principle of ensemble methods is to construct a linear combination of some model fitting method, instead of using a single fit of the method" (2012: 985). Here, we study the linear pool of *assessors* in the algorithmic fairness context.

Let $h_1, \ldots, h_m$ be a profile of m assessors on N. We focus on the well-behaved subset of linear pools that are convex combinations. For $\alpha_1, \ldots, \alpha_m \in [0, 1]$ such that $\sum_{j=1}^{m} \alpha_j = 1$, the linear pool h (without subscript) of the h_j for each $i \in N$ is given by

$$h(i) = \sum_{j=1}^{m} \alpha_j h_j(i). \qquad \text{(Linear Pool)}$$

Linear pooling allows us to weight the base assessors h_j the same or very differently. For instance, a weight of 0 might be assigned to models that were generated (e.g., by bagging) but were "pruned" from the final, pooled assessment (Mendes-Moreira et al., 2012: 3–4).

In general, the base assessors may fail to satisfy some criterion of fairness. But *when they do*—via a preprocessing, in-training, or post-processing approach (Barocas et al., 2023: pp. 70–71)—a crucial question is whether linear pooling preserves fairness criteria of the sort listed above since this would be a desirable feature whenever fairness is a concern. That is, given that each of $h_1, \ldots, h_m$ satisfy some criterion of algorithmic fairness, does it follow that the ensemble assessor h does? It depends on the criterion.

Theorem 1. *(i) Linear pooling* ***preserves*** *Base Rate Tracking, Ratio Base Rate Tracking, Equalized Odds, and Statistical Parity. (ii) Linear pooling does* ***not*** *preserve Calibration, Predictive Equity, Spanning, Strong Equalized Odds, Threshold Equalized Odds, or Strong Statistical Parity.*

While we do not provide counterexamples for criteria that are not preserved—none require big or complicated examples—proofs for the criteria that are preserved are in the Appendix.

There is a positive lesson and a negative lesson that we learn from Theorem 1. The positive lesson is that, for several conceptions of fairness, the linear pool of fair assessors is guaranteed to be fair. The set of properties preserved includes criteria from each of the three categories we mentioned: those related to Calibration, those related to Equalized Odds, and those related to Statistical Parity. The negative lesson is that, for other conceptions of fairness, the linear pool may not be fair *even if all of the* h_j *are*. The set of properties that fail to be preserved also includes criteria from each of the three categories. It may be worth remarking that the positive cases are criteria formulated in terms of *expectations* and the negative cases are criteria not formulated in those terms. This makes some sense mathematically since the central moves in the proofs exploit the linearity of expectation.

Alternatives to the linear pool have been studied in the pooling literature. Most prominent among these alternatives is probably geometric pooling. The most salient virtues of geometric pooling concern the various ways it interacts nicely with Bayesian updating (Madansky, 1964; Dietrich, 2019, 2021; Baccelli and Stewart, 2021). Again assuming that $\alpha_1, \ldots, \alpha_m \in [0, 1]$ and $\sum_{j=1}^{m} \alpha_j = 1$, the geometric pool is given by

$$h(i) = c \prod_{j=1}^{m} h_j(i)^{\alpha_j} \qquad \text{(Geometric Pool)}$$

where

$$c = \frac{1}{\prod_{j=1}^{m} h_j(i)^{\alpha_j} + \prod_{j=1}^{m} (1 - h_j(i))^{\alpha_j}}.$$

A less prominent alternative, closely related to geometric pooling, that has gained traction recently is the multiplicative pool (Dietrich, 2010; Easwaran et al., 2016). The multiplicative pool can be obtained from the geometric pool by setting $\alpha_j = 1$ for $j = 1, \ldots, m$.

$$h(i) = c \prod_{j=1}^{m} h_j(i), \qquad \text{(Multiplicative Pool)}$$

where

$$c = \frac{1}{\prod_{j=1}^{m} h_j(i) + \prod_{j=1}^{m} (1 - h_j(i))}.$$

While many things might be said in favor of these methods of pooling, that they preserve any of the fairness criteria discussed here is not one of them.

Theorem 2. *Neither Geometric nor Multiplicative pooling preserves Calibration, Base Rate Tracking, Ratio Base Rate Tracking, Predictive Equity, Spanning, Strong Equalized Odds, Equalized Odds, Threshold Equalized Odds, Strong Statistical Parity, or Statistical Parity.*

When preserving fairness is valuable, linear pooling has a definite advantage over geometric and multiplicative for many central statistical criteria. Theorems 1 and 2 are stated for cases in which every assessor in the profile satisfies some criterion. The issue can be studied from a more general perspective of *proximity* to satisfying a criterion rather than satisfying a criterion exactly. For example, is the linear pool of assessors guaranteed to be "at least as close" to satisfying some criterion as the base assessors on average? Addressing this more general perspective is beyond the scope of the current paper but is the subject of work in progress. But it is because of Theorems 1 and 2 that we focus on linear pooling in Section 4.

4 Accuracy and the Wisdom of Crowds

Recall that a primary motivation for using ensemble methods is improving the accuracy of forecasts. As we will show here by appealing to ideas from work on the wisdom of crowds, linear pooling does have potential accuracy benefits. *The wisdom of crowds* refers to the phenomenon of groups of individuals outperforming or being smarter than individuals in some sense (Galton, 1907; Surowiecki, 2004). Attempts to account for the phenomenon or its logic sometimes appeal to analytic results such as Condorcet's Jury Theorem (Condorcet, 1785) and the Diversity Prediction Theorem (Hong and Page, 2012). In this section, we note an analogue of the Crowds Beat Averages Law, a corollary of the Diversity Prediction Theorem, for assessors.

Define the inaccuracy of an assessor h on N as h's mean squared error:

$$\frac{1}{n} \sum_{i \in N} (h(i) - Y(i))^2.$$

Notice that implicit in this particular way of assessing inaccuracy is equal regard for the accuracy of forecasts for each individual. The inaccuracies of h for any two individuals are weighted the same—namely, $1/n$—in the overall inaccuracy of h.[1]

The next result is a Crowd Beats Averages Law for ensemble assessors.

Theorem 3. *Let h be a linear pool of a profile of m assessors $h_1, \ldots, h_m$ on N (such that the weights are positive and sum to 1). Then, the inaccuracy of h is no greater than the average inaccuracy of the h_j:*

$$\frac{1}{n}\sum_{i\in N}(h(i) - Y(i))^2 \leq \sum_{j=1}^{m} \alpha_j \frac{1}{n}\sum_{i\in N}(h_j(i) - Y(i))^2.$$

We can state Theorem 3 in more slogan form.

Ensemble Assessor Inaccuracy $\leq$ *Average Individual Assessor Inaccuracy*

It does not follow from Theorem 3 that the ensemble assessor is at least as accurate as *every* individual assessor in the profile, just that the ensemble's accuracy will be at least as good as the average accuracy of the assessors in the pool. And notice that the average on the right hand side is taken with respect to the pooling weights that determine h, where those weights need not be equal.

What Theorems 1.i and 3 help to make clear though is that linear pooling presents at once the prospect of improvements in accuracy *and* the aforementioned preservation of important fairness properties—a package that cannot be claimed for geometric or multiplicative pooling.

5 Discussion

To select from the "Pareto frontier" a class of assessors that are an optimal mix of accurate and fair requires a precise way of weighing those goals against each other (Kearns and Roth, 2019). We do not adopt this perspective here. But there is another, informal perspective that finds that imposing fairness constraints very often requires sacrifices in accuracy, and that gains in accuracy come at the cost of fairness. Theorems 1 and 3 bear on this informal notion of the accuracy-fairness tradeoff. In the context of assessors that take any value in the unit interval, we can exploit the propensity of ensemble methods to improve accuracy without sacrificing fairness for a number of statistical criteria studied in the literature, bucking the trend of finding accuracy-fairness tradeoffs. We basically have two guarantees. First, for the criteria indicated in Theorem 1.i, linear pools preserve fairness. Second, the inaccuracy of a linear pool is no greater than the average inaccuracy of individual assessors in the profile. So there are potential gains in accuracy that do not exact a fairness fee.

There are various lessons one could draw from our observations. Depending on one's antecedent sympathies for and commitments regarding the various fairness criteria, these points might be motivation for the use of ensemble methods in prediction tasks where fairness is a concern. This motivation, however, is limited in two respects. First, as far as our observations go, it is restricted to *linear* pooling. Second, it is limited to only four of the

[1] Our observations can be generalized to other ways of assessing inaccuracy (see, e.g., Pettigrew, MS).

ten criteria we present. A commitment to some fairness criterion that no prominent pooling method preserves may bring the accuracy-fairness tradeoff back in full force. Alternatively, if one is sold on ensemble methods and is concerned with fairness in some predictive setting, our points might be considerations in favor of particular statistical criteria. This can be only one consideration among others for favoring certain criteria. That a certain criterion is preserved by an ensemble assessor cannot be sufficient motivation for imposing the criterion since impossibility results have been established for some subsets of these very properties (Stewart et al., 2024).[2]

Funding

Lee Elkin was financially supported by the NWO-funded ENCODE project (VI.Vidi.191.105) and the Lamarr Institute for Machine Learning and Artificial Intelligence.

Appendix

Proof of Theorem 1

Proof. Let $h_1, \ldots, h_m$ be assessors on N, and let π be a partition of N. Let h be a linear opinion pool of $h_1, \ldots, h_m$. We establish each positive claim of the theorem in turn, omitting the counterexamples for the negative cases.

Base Rate Tracking. Suppose that h_j satisfies Base Rate Tracking for $j = 1, \ldots, m$. So for each j and for all $G, G' \in \pi$,

$$E_G(h_j) - E_{G'}(h_j) = \mu_G - \mu_{G'}.$$

Using this assumption and the linearity of expectation, we have

$$\begin{aligned}
E_G(h) - E_{G'}(h) &= E_G\left(\sum_{j=1}^{m} \alpha_j h_j\right) - E_{G'}\left(\sum_{j=1}^{m} \alpha_j h_j\right) \\
&= \sum_{j=1}^{m} \alpha_j E_G(h_j) - \sum_{j=1}^{m} \alpha_j E_{G'}(h_j) \\
&= \sum_{j=1}^{m} \alpha_j (E_G(h_j) - E_{G'}(h_j)) \\
&= \sum_{j=1}^{m} \alpha_j (\mu_G - \mu_{G'}) \\
&= \mu_G - \mu_{G'}.
\end{aligned}$$

[2]Thanks to Jean Baccelli, Reuben Stern, Michael Nielsen, and two anonymous referees for helpful feedback. We would also like to thank Frederik Van De Putte, Nicolien Janssens, Eva Schmidt, and Jakob Rehof for helpful comments on a separate project led by Lee Elkin that merged with this one.

Hence, h, an arbitrary linear pool of $h_1, \ldots, h_m$, satisfies Base Rate Tracking.

Ratio Base Rate Tracking. Suppose that h_j satisfies Ratio Base Rate Tracking for $j = 1, \ldots, m$. So for all j and all $G, G' \in \pi$,

$$\frac{E_G(h)}{\mu_G} = \frac{E_{G'}(h)}{\mu_{G'}}.$$

Using this assumption and the linearity of expectation,

$$\begin{aligned}
\frac{E_G(h)}{\mu_G} &= \frac{E_G(\sum_{j=1}^m \alpha_j h_j)}{\mu_G} \\
&= \frac{\sum_{j=1}^m \alpha_j E_G(h_j)}{\mu_G} \\
&= \sum_{j=1}^m \alpha_j \frac{E_G(h_j)}{\mu_G} \\
&= \sum_{j=1}^m \alpha_j \frac{E_{G'}(h_j)}{\mu_{G'}} \\
&= \frac{E_{G'}(h)}{\mu_{G'}}
\end{aligned}$$

Hence, h, an arbitrary linear pool of $h_1, \ldots, h_m$, satisfies Ratio Base Rate Tracking.

Equalized Odds. Suppose that h_j satisfies Equalized Odds for $j = 1, \ldots, m$. So for all j and for all $G, G' \in \pi$,

$$E_G(h_j|Y = 0) = E_{G'}(h_j|Y = 0)$$

and

$$E_G(1 - h_j|Y = 1) = E_{G'}(1 - h_j|Y = 1).$$

Consider equal generalized false positive rates first. Using the assumption and the linearity of expectation,

$$\begin{aligned}
E_G(h|Y = 0) &= E_G\left(\sum_{j=1}^m \alpha_j h_j|Y = 0\right) \\
&= \sum_{j=1}^m \alpha_j E_G(h_j|Y = 0) \\
&= \sum_{j=1}^m \alpha_j E_{G'}(h_j|Y = 0) \\
&= E_{G'}\left(\sum_{j=1}^m \alpha_j h_j|Y = 0\right) \\
&= E_{G'}(h|Y = 0)
\end{aligned}$$

An analogous argument establishes that $E_G(1-h|Y=1) = E_{G'}(1-h|Y=1)$. Hence, h, an arbitrary linear pool of $h_1, \ldots, h_m$, satisfies Equalized Odds.

Statistical Parity. This proof is closely analogous to the proof for equal generalized false positive rates, replacing the conditional expectations $E_G(h|Y=0)$ and $E_{G'}(h|Y=0)$ with expectations conditional on only group membership: $E_G(h)$ and $E_{G'}(h)$.

□

Proof of Theorem 3

Proof. Note first that, for each $i \in N$,

$$(h(i) - Y(i))^2 \leq \sum_{j=1}^{m} \alpha_j (h_j(i) - Y(i))^2. \tag{1}$$

Equation 1 is a slightly generalized version of the Crowds Beat Averages Law for point estimates, and follows immediately by Jensen's Inequality. From 1, it follows that

$$\frac{1}{n}\sum_{i \in N}(h(i) - Y(i))^2 \leq \frac{1}{n}\sum_{i \in N}\sum_{j=1}^{m} \alpha_j (h_j(i) - Y(i))^2. \tag{2}$$

Finally, switching the order of summation and rearranging on the right hand side of 2 gives the desired result:

$$\frac{1}{n}\sum_{i \in N}(h(i) - Y(i))^2 \leq \sum_{j=1}^{m} \alpha_j \frac{1}{n}\sum_{i \in N}(h_j(i) - Y(i))^2.$$

□

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016a, May). How we analized the compas recidivism algorithm. `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016b). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Baccelli, J. and R. T. Stewart (2021). Support for geometric pooling. *The Review of Symbolic Logic 16*(1), 298–337.

Barocas, S., M. Hardt, and A. Narayanan (2023). *Fairness and Machine Learning*. Cambridge, MA: MIT Press. `http://www.fairmlbook.org`.

Borsboom, D., J.-W. Romeijn, and J. M. Wicherts (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods 13*(2), 75–98.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In J. Gentle, W. Härdle, and Y. Mori (Eds.), *Handbook of Computational Statistics: Concepts and Methods*, pp. 985–1022. Berlin, Heidelberg: Springer.

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

Condorcet, M. d. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. The Royal Printing Office.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM.

Dietrich, F. (2010). Bayesian group belief. *Social Choice and Welfare 35*(4), 595–626.

Dietrich, F. (2019). A theory of bayesian groups. *Noûs 53*(3), 708–736.

Dietrich, F. (2021). Fully bayesian aggregation. *Journal of Economic Theory 194*, 105255.

Dietrich, F. and C. List (2016). Probabilistic opinion pooling. In A. Hájek and C. Hitchcock (Eds.), *Oxford Handbook of Probability and Philosophy*. Oxford University Press.

Easwaran, K., L. Fenton-Glynn, C. Hitchcock, and J. D. Velasco (2016, June). Updating on the credences of others: Disagreement, agreement, and synergy. *Philosophers' Imprint 16*(11), 1–39.

Elkin, L. and R. Pettigrew (2025). *Opinion Pooling*. Elements in Decision Theory and Philosophy. Cambridge University Press.

Eva, B. (2022). Algorithmic fairness and base rate tracking. *Philosophy & Public Affairs 50*(2), 239–266.

Galton, F. (1907). Vox populi. *Nature 75*, 450–451.

Genest, C. and J. V. Zidek (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science 1*(1), 114–135.

Gilboa, I., D. Samet, and D. Schmeidler (2004). Utilitarian aggregation of beliefs and tastes. *Journal of Political Economy 112*(4), 932–938.

Grant, D. G. (2023). Equalized odds is a requirement of algorithmic fairness. *Synthese 201*(3), 101.

Grgić-Hlača, N., M. B. Zafar, K. P. Gummadi, and A. Weller (2017). On fairness, diversity and randomness in algorithmic decision making. *arXiv preprint arXiv:1706.10208*.

Hong, L. and S. E. Page (2012). Some microfoundations of collective wisdom. In H. Landemore and J. Elster (Eds.), *Collective Wisdom: Principles and Mechanisms*, pp. 56–71. Cambridge University Press.

Kearns, M. and A. Roth (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Leutbecher, M. and T. N. Palmer (2008). Ensemble forecasting. *Journal of Computational Physics 227*(7), 3515–3539.

Maclin, R. and D. Opitz (1997). An empirical evaluation of bagging and boosting. *AAAI/IAAI 1997*, 546–551.

Madansky, A. (1964). Externally bayesian groups. Santa Monica, CA: RAND Corporation.

McConway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association 76*(374), 410–414.

Mendes-Moreira, J., C. Soares, A. M. Jorge, and J. F. D. Sousa (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys 45*(1), 1–40.

Menon, A. K. and R. C. Williamson (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118.

Mongin, P. (1995). Consistent bayesian aggregation. *Journal of Economic Theory 66*(2), 313–351.

Narayanan, A. (2018). 21 fairness definitions and their politics (tutorial). *Conference on Fairness, Accountability & Transparency*, `https://www.youtube.com/watch?v=jIXIuYdnyyk`.

Nielsen, M. (2019). On linear aggregation of infinitely many finitely additive probability measures. *Theory and Decision 86*(3-4), 421–436.

Nielsen, M. and R. T. Stewart (2024). New possibilities for algorithmic fairness. *Philosophy & Technology 37*(116).

Pettigrew, R. (2019). *On the Accuracy of Group Credences*, Volume 6, Chapter 6, pp. 137–160. Oxford: Oxford University Press.

Pettigrew, R. (MS). Generalizing the diversity prediction theorem. Unpublished Manuscript.

Räz, T. (2021). Group fairness: Independence revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 129–137.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review 33*, 1–39.

Stewart, R., B. Eva, S. Slank, and R. Stern (2024). An impossibility theorem for base rate tracking and equalised odds. *Analysis*, Forthcoming.

Stewart, R. T. (2024). The ideals program in algorithmic fairness. *AI & Society*, Forthcoming.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter than the Few*. London: Anchor.

Verma, S. and J. Rubin (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pp. 1–7.