Algorithmic Fairness: A Tale of Two Approaches

Joshua S. Gans

December 3, 2024

Abstract

The growing use of artificial intelligence in high-stakes decision-making has raised important questions about how to address potential discriminatory outcomes. Two distinct approaches have emerged: one from computer science, which focuses on regulating algorithms directly, and another from economics, which emphasises market design and incentives. This paper examines these competing frameworks and argues that the economic approach leads to superior welfare outcomes.

1 Introduction

As algorithms based on artificial intelligence proliferate, an important regulatory challenge is whether and how to regulate their use to meet various goals associated with achieving fairness and/or preventing discrimination. To date, policy-makers have focussed on computer science-led approaches to these issues. However, there is an emerging literature from economists that challenges this approach to regulation.

The computer science approach, as outlined by Barocas et al. (2023), focuses on imposing fairness constraints directly on algorithmic systems. This typically involves mathematical definitions of fairness - such as demographic parity, equal opportunity, or equal odds - that algorithms must satisfy. For example, demographic parity requires that predictions be independent of protected group status, while equal opportunity mandates equal true positive rates across groups. These constraints are then incorporated into the algorithm's training process. Thus, algorithms are regulated directly by altering the types of predictions such algorithms can deliver.

More recently, economists have investigated this issue using standard approaches to devising regulatory outcomes based on decision-making contexts. As Kleinberg et al. (2016) demonstrates, the various computer science-developed fairness criteria are generally incompatible with each other except in trivial cases. More importantly, Rambachan et al. (2020)

show that directly regulating algorithms by removing protected attributes or enforcing fairness constraints can harm disadvantaged groups by reducing prediction accuracy in ways that disproportionately affect them. Thus. instead of regulating prediction outcomes, the economics approach identifies how algorithmic predictions are used. It focuses on the broader decision-making context rather than the algorithms themselves. As articulated by Rambachan et al. (2020), the key insight is that AI predictions should be allowed to use all available information to maximise accuracy, while fairness objectives should be achieved through group-specific decision thresholds. Their theoretical analysis demonstrates that this approach dominates direct algorithmic regulation in terms of both efficiency and equity.

2 A Formal Analysis of Approaches

To understand the fundamental difference between computer science and economic approaches to algorithmic fairness, it is helpful to express them formally. Let us consider a setting where a decision-maker must allocate K slots among candidates from two groups ($g \in \{0, 1\}$), where group 0 represents a historically disadvantaged group.

The standard computer science approach typically starts with an optimisation problem of the form:

$$\min_f \frac{1}{N} \sum L(\omega_i, X_i; f) \text{ subject to } f \text{ being `fair'} \tag{1}$$

where f is the algorithm, L is a loss function, ω_i is the true outcome of interest, and X_i represents observable characteristics for each $i \in \{1, ..., N\}$. The fairness constraint might take various forms, such as demographic parity:

$$\Pr[s_i = 1 | g = 1] = \Pr[s_i = 1 | g = 0] \tag{2}$$

where s_i is the algorithm's prediction consistent with being allocated to a slot. This approach directly constrains the algorithm's outputs to achieve fairness goals.

In contrast, the economic approach articulated by Rambachan et al. (2020) formulates the problem as:

$$\max_{a(g,x)} \mathbb{E}\left[\sum_{g,x} \phi_g \omega(g,x) a(g,x) F(g,x)\right] \text{ subject to } \sum_{g,x} a(g,x) F(g,x) \leq K \tag{3}$$

where $a(g, x) \in \{0, 1\}$ is the allocation decision, ϕ_g represents welfare weights that can differ across groups, and $F(g, x)$ is the distribution of characteristics. Crucially, this formulation separates the prediction problem from the allocation decision with ϕ_g being a measure of what Agrawal et al. (2018) refer to as *judgment* (in this case, of the regulator or planner).

The solution takes the form of group-specific thresholds $t^*(g)$, where a candidate is selected if:

$$\mathbb{E}[\omega_i(g_i, x_i)] > t^*(g)$$

The economic formulation has several advantages. First, it allows the algorithm f to use all available information to maximise prediction accuracy, which benefits all groups. Second, it explicitly incorporates social preferences through welfare weights ϕ_g rather than through algorithmic constraints. Finally, it recognises that optimal policy may involve different treatment thresholds across groups even when the underlying predictions are unbiased.

To see why this matters quantitatively, note that:

$$s_i = \underbrace{\omega_i}_{\text{Outcome of Interest}} + \underbrace{\Delta_{\omega_i}}_{\text{Measurement Error}}$$

The difference in the algorithm's estimated predictions, $\mathbb{E}[s_i|g=1] - \mathbb{E}[s_i|g=0]$, is decomposed by Rambachan et al. (2020) as:

$$\underbrace{\mathbb{E}[\omega_i|g=1] - \mathbb{E}[\omega_i|g=0]}_{\text{base rate differences}} + \underbrace{\mathbb{E}[\Delta_{\omega_i}|g=1] - \mathbb{E}[\Delta_{\omega_i}|g=0]}_{\text{measurement error differences}} + \underbrace{\hat{\epsilon}(1) - \hat{\epsilon}(0)}_{\text{estimation error differences}}$$

where $\hat{\epsilon}(g) = \mathbb{E}[\theta_i|g] - \mathbb{E}[s_i|g]$. This shows that there are three broad reasons why a signal might differ from the outcome of interest. First, there are estimation errors that arise across groups and may come from differences in data availability or precision. Second, there are measurement errors that arise because there may be model misspecification in the training data or bias if how the training data is collected. Finally, there are base rate differences that are real differences between groups in terms of expected performance that arise because of underlying socio-economic or other reasons. The important thing to recognise is that while the first two reasons can be corrected by improvements in the algorithm, data or an understanding of the source of measurement issues, the final reason cannot. Importantly, the computer science approach attempts to eliminate this entire difference through algorithmic constraints. In contrast, the economic approach addresses only the latter two terms through improved prediction while handling base rate differences through decision thresholds.

The intuition favouring the economic approach is straightforward - accurate predictions allow decision-makers to better identify qualified candidates from disadvantaged groups who might otherwise be overlooked. The decision thresholds can then be adjusted to achieve desired representation goals while maintaining the benefits of accurate prediction. This first-best solution recognises that statistical discrimination arising from genuine differences in base rates cannot be eliminated through algorithmic constraints alone.

A concrete example helps illustrate this distinction. Consider college admissions, where an AI system predicts student success based on test scores and other factors. The computer science approach might require equal admission rates across groups by constraining the algorithm. The economic approach would instead let the algorithm make the most accurate predictions possible, while potentially setting different admission thresholds for different groups based on broader social welfare considerations.

This distinction becomes particularly important when considering welfare implications. Under reasonable conditions, Rambachan et al. (2020) show that the economic approach weakly dominates any algorithm satisfying typical fairness constraints. The intuition is that accuracy-maximizing prediction combined with optimal threshold policies can always replicate the outcome of a constrained algorithm, but may be able to achieve better outcomes by utilising information more efficiently.

This formal analysis thus provides strong theoretical and practical arguments for favouring the economic approach. While both frameworks share the goal of reducing discriminatory outcomes, the economic approach achieves this more efficiently by preserving prediction accuracy while using decision thresholds (that are based directly on judgment) to achieve distributional objectives.

The economic framework also better accounts for the reality that decision-makers may have legitimate preferences for diversity or equity that go beyond pure prediction. Rather than conflating these preferences with the prediction task itself, it provides a cleaner separation that allows for more transparent policy choices.

None of this is to suggest that algorithmic bias is not a serious concern requiring policy intervention. Indeed, the economic approach may ultimately require more active regulation of human decision-makers rather than algorithms.[1] However, it suggests that attempting to solve discrimination through direct algorithmic regulation is likely to be both ineffective and counterproductive.

3 The Fairness Frontier Framework

The above analysis highlights the distinction between regulating prediction algorithms themselves versus regulating how predictions are used. There is another approach to this trade-off that takes the perspective that different algorithms themselves involve a trade-off between accuracy and fairness. The underlying motivation for this approach is that in many regu-

[1] This is discussed in detail in Gans (2025). Chapter 20.

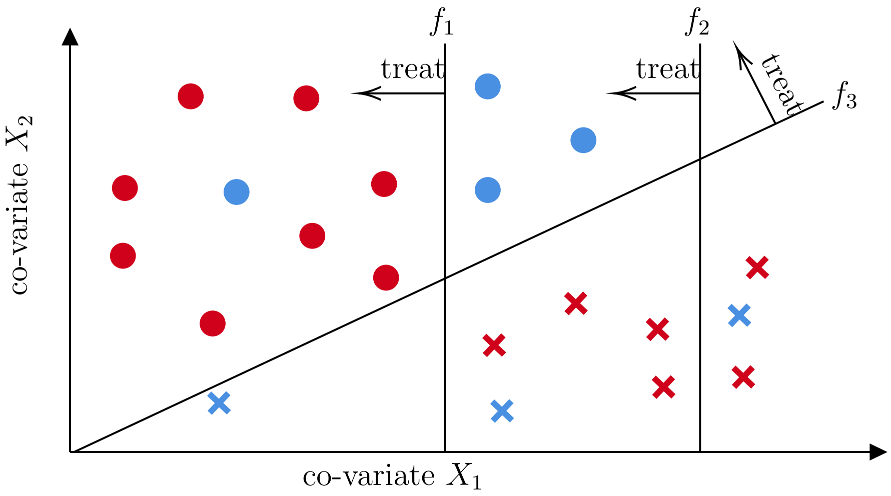


Figure 1: Fairness and Accuracy of Different Algorithms

latory contexts where discrimination or unfairness is regulated, a potential defence is that, in so doing, there would be a cost in terms of accuracy. However, it could also be the case that for a particular algorithm, no such trade-off arises. Identifying such cases, however, requires both understanding the particular fairness criterion applied and understanding the risk preferences of the social planner.

This approach is outlined in Liang et al. (2024). The idea that there is a trade-off between fairness and accuracy is easy to assess. Consider the distribution of outcomes in Figure 1. There, X is comprised of two covariates (e.g., random variables), X_1 and X_2, that inform whether a patient should be treated or not. The circular points are helped by the medical treatment, and the crosses are harmed by it. There are various algorithms indicated by demarcation lines indicating who would be treated or not. Algorithms f_1 and f_2 use only X_1 as an input, while f_3 uses both covariates. It can seen in f_1 that no red patient is harmed while 4/7 of the blue patients are harmed. Overall 1/21 receives unnecessary treatment while 3/21 miss out on treatment they need. For f_2, the error rates are equalised at a rate of 2/7 for each group, demonstrating a fairness-accuracy tradeoff as overall 9/21 are harmed, all receiving unnecessary treatment. Notice that if everyone is treated, in this example, the error rates are also equalised at 3/7 for each group. Finally, for f_3, both co-variates are available, and there are no errors. Thus, adding a co-variate can improve accuracy and fairness, and this co-variate could be identifiable as group membership.

Building on our previous formal analysis, the fairness frontier approach developed by Liang et al. (2024) provides a powerful framework for understanding the trade-offs inherent in algorithmic fairness and offers insights into optimal policy design. Let me explain this approach formally and demonstrate why it leads to superior outcomes.

Consider a decision-maker whose utility from algorithm predictions depends on two types of errors for each group g:

$$e_g = (1 - \Pr[s_i = 1|\omega_i = 1, g])\ell_1 + (1 - \Pr[s_i = 0|\omega_i = 0, g])\ell_0 \tag{4}$$

where ℓ_1 and ℓ_0 represent the losses from false negatives and false positives respectively. The fairness frontier can then be characterized in $\{e_0, e_1\}$ space - that is, the error rates for groups 0 and 1.

A key innovation in their framework is the definition of algorithmic dominance. Algorithm f is said to strictly dominate algorithm f' if:

$$e_0(f) \leq e_0(f') \text{ and } e_1(f) \leq e_1(f') \tag{5}$$

with at least one strict inequality. This allows us to identify Pareto optimal algorithms that cannot be improved for one group without harming the other.

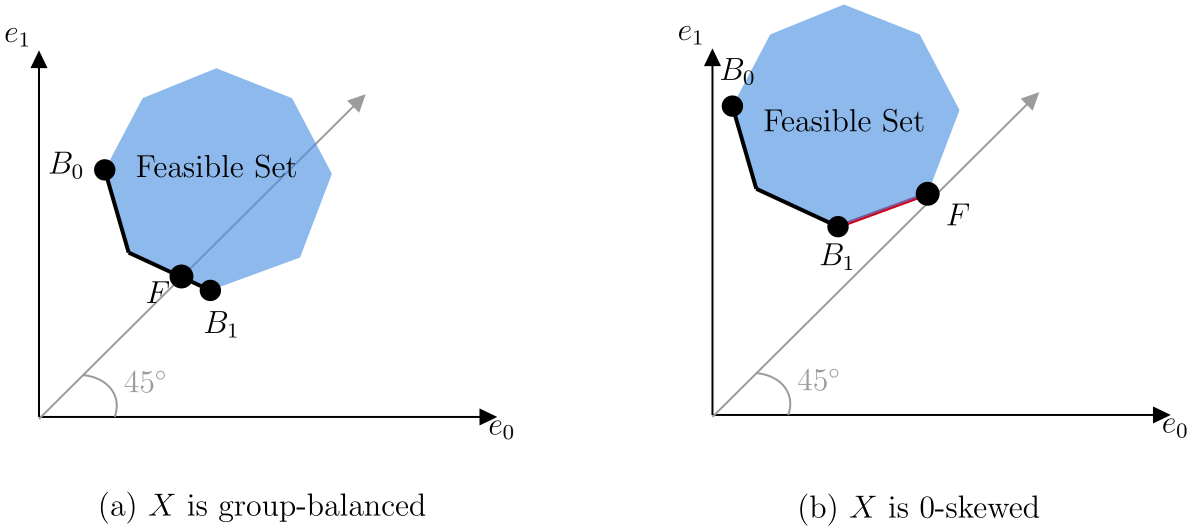
The fairness frontier consists of algorithms that minimise a weighted average of group errors:

$$\min_f \lambda e_0(f) + (1 - \lambda)e_1(f) \tag{6}$$

for different values of $\lambda \in [0, 1]$. This generates a curve in error space that represents the fundamental trade-offs between accuracy for different groups.

Figure 2 provides a graph of the fairness-accuracy frontier on $\{e_1, e_0\}$ space. Two possibilities are depicted. In each, there are three points of interest. The first two are algorithmic choices that result in the best outcomes for members of a particular group g, B_1 and B_0, respectively, minimising amongst feasible outcomes e_1 and e_0, respectively. The final one, F, is the one that is the most fair and minimises $|e_1 - e_0|$. Note that in both, the accuracy frontier lies on the lower portion of the feasible set between B_1 and B_0. The point that is most fair, F, lies on that frontier in (a) but not in (b). This means that a planner can choose a fair algorithm that does not sacrifice accuracy in (a) but faces a trade-off in (b).

Liang et al. (2024) show that the difference between the two types of situations depends on the properties of X, the set of observable variables other than group membership that is used to generate predictions. X is group-balanced if an algorithm that is optimal for a group (say, B_1) gives a lower error for $g = 1$ than for $g = 0$. If this condition does not hold, then X is group-skewed. In Figure 2, X is 0-skewed and so has a lower error for group 0 and both B_1 and B_0 than the respective errors at those points for group 1. Group balance arises when X has different implications for each group – e.g., a high X implies a high ω for 0 and a lower ω for 1. This could also arise if different dimensions of X are uninformative about ω



(a) X is group-balanced

(b) X is 0-skewed

Figure 2: Fairness-Accuracy Frontier

for each group. Group skewness arises when X is asymmetrically informative. For instance, when medical data is recorded more accurately for high-income patients than low-income patients, making the prediction more dispersed for high-income patients than low-income ones.[2]

This analysis assumes that the planner chooses the algorithm. In some situations, as we have noted earlier, the regulator chooses the inputs that can be used to generate predictions. Liang et al. (2024) shows that this makes the problem one of information design where algorithm designers (choosing algorithms for their own goals) use data that is garbled by the regulator (i.e., choosing a coarser partition of the data or leaving out certain variables or introducing an error for, say, privacy reasons). Such garblings can potentially change the feasible set of algorithms and push the agent towards the planner's preferred algorithm. This provides another perspective on whether to, say, ban the use of test scores in admission decisions. Those scores are likely to be decision-relevant for both groups. Thus, if group identity, g, is available, excluding test scores will reduce welfare when algorithm designers can garble components of X. By contrast, if g is not available, a fairness-motivated algorithm designer might choose to exclude test scores.

Once again, this framework demonstrates why computer science approaches focusing solely on fairness constraints can be suboptimal. Consider a typical fairness constraint like:

$$|e_0 - e_1| \leq \epsilon$$

[2] Particular special cases can refine this analysis further. For instance, when g is an informative part of the input data, then $B_1 = B_0$ and the fair outcome potentially is Pareto optimal; if $\omega|X$ is independent of g, then once you have X you do not have any additional predictive value from g and if there is strong independence in that (X, ω) is independent of g then $B_1 = B_0 = F$.

This constraint may force the algorithm to operate at point F even when there exist points on the frontier that would improve outcomes for both groups. In contrast, the economic approach allows movement along the frontier to find optimal points based on social welfare weights.

Formally, the optimal policy solves:

$$\max_{f,t_0,t_1} W = \sum_g \phi_g[u(\omega_g, a_g)] \text{ subject to } a_g = 1 \text{ if } f(x_g) > t(g) \tag{7}$$

where t_g are group-specific thresholds and ϕ_g are welfare weights.

This formulation allows us to:

1. Choose an algorithm on the efficiency frontier
2. Set thresholds to achieve distributional objectives
3. Explicitly account for social preferences

Using real-world data, Liang et al. (2024) shows that many commonly used fairness constraints force algorithms to operate well inside the frontier, creating unnecessary inefficiency. Their analysis suggests potential Pareto improvements of 10-30% in both accuracy and fairness metrics by moving to frontier-optimal algorithms with appropriate thresholds.

This framework provides a geometric interpretation of the decomposition, $\mathbb{E}[s_i|g = 1] - \mathbb{E}[s_i|g = 0] = \text{Base Rate} + \text{Measurement} + \text{Estimation}$, we discussed earlier. The shape of the frontier reflects these components, with base rate differences creating fundamental trade-offs while measurement and estimation errors can potentially be eliminated through better algorithm design.

The frontier framework suggests a two-step approach to policy, First, identify the achievable frontier through empirical analysis. Second, choose operating points based on explicit welfare criteria. This approach dominates direct fairness constraints because it maintains maximum predictive accuracy, makes trade-offs explicit, allows for welfare-optimal solutions and identifies achievable improvements.

The framework thus provides both theoretical justification and practical guidance for favouring economic approaches to algorithmic fairness over computer science constraints. It demonstrates that many apparent accuracy-fairness trade-offs are artificial products of suboptimal regulation rather than fundamental limitations.

4 Conclusion

The economic framework provides both theoretical justification and practical guidance for favouring economic approaches to algorithmic fairness over computer science constraints. It demonstrates that many apparent accuracy-fairness trade-offs are artificial products of suboptimal regulation rather than fundamental limitations.

This perspective aligns with broader economic insights about the importance of getting incentives right rather than trying to directly control outcomes. Just as price controls are generally inferior to targeted subsidies for achieving distributional goals, algorithmic fairness constraints may be inferior to properly structured decision rules and thresholds.

The key policy implication is that regulators should focus less on constraining algorithms themselves and more on the broader institutional context in which they are deployed. This might include requirements for transparency about decision criteria, audits of outcome disparities, and guidelines for appropriate use of group-specific thresholds. Such an approach would better balance the competing goals of accuracy and fairness while avoiding the pitfalls of direct algorithmic regulation.

References

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review Press, 2018.

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

Gans, Joshua S., *The Microeconomics of Artificial Intelligence*, MIT Press, 2025.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

Liang, Annie, Jay Lu, Xiaosheng Mu, and Keiji Okumura, "Algorithm design: A fairness-accuracy frontier," Technical report 2024.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan, "An economic perspective on algorithmic fairness," *AEA Papers and Proceedings*, 2020, *110*, 91–95.