Algorithmic Fairness: Some Practical Considerations for Actuaries

Prepared by Chris Dolman and Dimitri Semenovich

Presented to the Actuaries Institute
Actuaries Summit
3 – 4 June 2019

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Algorithmic Fairness: Some Practical Considerations for Actuaries

Chris Dolman, Dimitri Semenovich

Abstract: In this paper we review recent work on fairness of decision making algorithms and highlight some important relationships between recent proposals from the computer science literature and traditional concepts such as actuarial fairness. We also explore some practicalities and implications of the use of fairness constructs within real world pricing systems to provide guidance for members who wish to adopt these ideas in their work.

Keywords: Algorithmic Fairness, Data Analytics, Insurance Pricing, Ethics, Model Risk

Acknowledgments: The authors would like to thank Bartosz Piwcewicz for reviewing this paper. We note this paper is an extension of our earlier work on this topic (Dolman and Semenovich, 2018a; Dolman and Semenovich, 2018b).

1 Background and Overview

As society has become increasingly dependent on personal data collection and automated decision making based on such data (commonly referred to as "algorithmic decisions"), the public concern about the potential for these practices to do harm has also grown.

One area in which algorithms may cause harm is in fostering discrimination or, more generally, a lack of fairness within decisions. Acknowledgement of this has given rise to a cross-disciplinary subfield of academic study, often referred to as "algorithmic fairness" or "quantitative fairness". Whilst much of the research is in its early stages we believe it should prove valuable to actuaries and data scientists involved in building or managing decision making algorithms, particularly as broader study directed towards the topic may require us to re-examine historic normative practices. Improving awareness of this research is our first objective.

Our second objective is to illustrate some potential opportunities for actuaries in the algorithmic fairness arena. We identify some inherent limitations of a purely data driven approach and highlight the critical importance of judgement and domain expertise, warranting some further attention from the actuarial profession.

Finally, we examine how some common insurance pricing practices may relate to the broader fairness research, establishing a linkage to the recent literature.

We hope to encourage more actuaries to contribute to the current societal debate around fairness of algorithms, an area in which we feel actuaries can provide a valuable perspective.

2 Motivating Example - The COMPAS debate

Much of the most recent academic research has been inspired by high profile examples of alleged unfairness of an algorithm. We will use one of them here as a motivational example.

In May 2016, the investigate journalism website Propublica issued a challenge of racism towards an algorithm used in the US criminal justice system (Angwin et al., 2016). The algorithm in question was called the "Correctional Offender Management Profiling for Alternative Sanctions", or COMPAS, produced by a company Northpointe, Inc.

The aim of the COMPAS algorithm is to produce an objective, data derived "risk score" for different forms of recidivism. Algorithms of this form are used in the criminal justice setting to help inform particular decisions such as granting of bail or parole. The final decision still sits with the judge – the risk score is intended to be used to inform, not to make, the final decision for discretionary programmes. The "risk score" uses a scale of $0 - 10$ to denote the risk of an individual re-offending, based on a model driven by many dozens of attributes. Notably, race is not one of those attributes.

The case against the use of the algorithm was multifaceted, but at its core was a simple observation based on data obtained from Broward County, Florida. Propublica noted that in the two years following the COMPAS score being applied, the rate of "false positives" varied by race:

- of those people who did not reoffend, those who were black were classified as "higher risk" at a rate of 45%,
- of those people who did not reoffend, those who were white were classified as "higher risk" at a rate of 23%.

Propublica's main conclusion was that this is unfair: of those who ultimately did not reoffend, blacks were far more likely than whites to have been rated as "high risk". Having been rated "high risk", people would likely have been subject to harsher decisions around parole, bail, etc., yet ultimately they were not in fact rearrested or convicted and so were perhaps unduly disadvantaged by being issued with this classification. Whilst some errors of this form are unavoidable, the errors were disproportionately affecting black people. Similarly, Propublica also found that for those people who ultimately did go on to reoffend, whites were substantially more likely than blacks to have been rated as "lower risk".

The maker of the algorithm, Northpointe, did not accept the charge of racism. They demonstrated that the predictive accuracy of their algorithm at each point of their risk score scale was effectively the same, irrespective of race (Dieterich et al., 2016). Put simply, a high risk score *meant the same thing*, in terms of the predicted chance of reoffending, no matter whether you were black or white. Indeed, Northpointe researchers had commented on model accuracy in relation to race (and gender) in previous publications (e.g. Brennan et al., 2008).

There then followed a flurry of interest in both the mathematics of the situation and the social policy side. The Washington Post published an excellent summary of the mathematical arguments (Corbett-Davies et al., 2016), and even the Wisconsin Supreme Court observed and commented on the debate (State of Wisconsin vs Loomis, 2016).

Our observations are as follows:

- At a high level, the points made by Propublica and Northpointe are reasonable observations of the actual data. There is clearly no substantive argument of fact. Instead, the debate seems to stem from a disagreement over the correct goal, or aim, of the algorithm, as it pertains to fairness within the justice system.
- Northpointe's position is that the result of an algorithm should mean the same thing irrespective of race. A score of 8 should represent the same piece of information (in this case risk of reoffending) regardless of race. It is certainly hard to suggest that this should not be the case – if instead we treated people with the same score differently based on race, or equivalently if the scores were differently calibrated for each racial group, racism could certainly also be alleged.
- Propublica's position is that the distribution of false positives is inequitable. They make a similar, separate point about false negatives. Again, it is hard to disagree with the point – certainly it seems intuitive to think that we should be concerned that black people who are not found to reoffend are more likely to be classed as high risk, given that this classification will tend to mean harsher treatment.

- As we will discuss later in this paper, for situations such as this it is unfortunately not possible for a classification algorithm to meet all three of these definitions of fairness simultaneously, except in narrow, trivial, circumstances (Kleinberg et al., 2017; Chouldechova, 2017). Tradeoffs are unavoidable.
- The conclusion about fairness metric incompatability was made possible due to the mathematisation of the problem; the formal approach has demonstrated that the real problem was one of definition, not necessarily intent. Certainly, we expect that nobody set out to be deliberately unfair or discriminatory, however, the adversarial nature of the public discussion was not conducive to a nuanced debate over trade-offs that the situation demanded. COMPAS is still being used today.

Cases such as COMPAS illustrate that even though there are societal norms, and indeed laws, against discrimination, by analysing the decision in a mathematical form we find that we may need to think more precisely about exactly what norms we are looking to encode and enforce. A mere commitment to "be fair" is perhaps not good enough.

3 Fairness Concepts and Metrics

In this section we review several of the more common fairness concepts and metrics from contemporary research, many of which were motivated by studies of individual algorithms such as COMPAS. In each case, we include mathematical definitions of the concept, followed by a discussion of the merits and drawbacks of it. Where possible, we draw analogies to insurance pricing practices. Many of the concepts are highly intuitive so the readers are encouraged not to dwell on the somewhat abstract technical defintions.

3.1 Notation

In all the examples that follow we shall use the following notation:

- Y represents the observed outcome of interest (for example, observed to reoffend, default on a loan, realised cost of claims or similar). Y can either take values in $\{0, 1\}$ in the binary classification case or in $\mathbb{R}$ more generally.
- A represents an observed protected attribute (for example, race, age, gender, etc). For ease, we will let A take two values, a and a', without substantive loss of generality in the points made.
- X represents an observed vector of attributes for each individual in the population, that are not protected and can be used for prediction tasks.
- μ represents the true "type" associated with the individual that is usually unobserved. For example, μ may represent whether the individual is intent on commiting fraud or reoffending. In the typical binary classification case it is often implicitly assumed that $\mu = Y$, i.e. that the type is always revealed through the outcome Y. In the insurance pricing case μ would represent the "true" expected cost of claims at the start of the contract.
- $d(X, A)$ represents a decision function (for example, to issue a loan or not, or the actual price in market for an insurance contract), taking values in $\{0, 1\}$ or $\mathbb{R}$ as appropriate. Decision d is commonly defined relative to a threshold or some other transformation over a model that has been constructed in an attempt to predict Y. Questions of fairness usually revolve around d: is the decision procedure fair, having regard to observable data?

3.2 Unawareness

This is the often the first approach taken in an attempt to create "fair" decision making algorithms. "Unawareness" requires us to not explicitly consider the protected attribute A in the

decision procedure d:

$$d(x, a) = d(x, a'), \quad \forall x \in X.$$

In some sense this parallels the concept of "direct discrimination" which is commonly legislated against. To the best of our knowledge this is the common approach taken to address the EU Gender Directive in the setting of insurance premiums – if we were to change someone's gender and leave all other inputs equal, the price does not change.

Frequently, where privacy legislation exists we cannot collect certain items of sensitive data. Hence we are often under a situation of "unawareness" by default – if the attribute of data cannot be collected, it cannot be used in the decision procedure. Whilst insurance is often a special case under privacy and discrimination legislation in many countries, allowing certain items of sensitive data such as age to be collected and used to price policies, there are still many attributes which may not be collected.

Owing to the well known problem of redundant encoding of protected attributes (for example, strong correlations between location and race in some areas) "unawareness" has been frequently criticised as vulnerable to indirect discrimination and disparate impact (for example, Pedreshi et al., 2008).

There is a natural conflict between complying with this ideal of unawareness (and hence concepts of direct discrimination), and complying with many of the notions that we shall come to below which are analogous to indirect discrimination. In particular, we note that many proposals for correction mechanisms to ensure algorithmic fairness use the protected attribute directly as part of a corrective process (for example the post-processing step of Hardt et al., 2016).

3.3 Demographic Parity

A perhaps more demanding fairness metric is "demographic parity" or "statistical parity". This concept requires expected value of the decision to be equal across protected classes. We can state it formally as:

$$\mathbb{E}(d \,|\, A = a) = \mathbb{E}(d \,|\, A = a'),$$

noting that for binary classification $\mathbb{E}(d) = \Pr(d = 1)$ or the probability of a positive decision. At one level this is appealing. If d is the decision of a bank to lend money, for example, and the protected class is gender, this metric requires the probability that a loan is given to be the same for each gender.

However, demographic parity may lead to complications. Again using loans as an example, if a particular protected group is generally not creditworthy and hence very prone to default, the use of a decision procedure complying with demographic parity might lead to greater harm to the protected group than other decision procedures, since defaulting on a loan may be a worse outcome than getting no loan at all. The idea that implementing fairness metrics may cause unintended harm, particularly over time, is something the research community is beginning to grapple with (Liu et al., 2018).

In other situations this concept is less problematic. For example, we might choose to use demographic parity for an advertising campaign for a job, to ensure that all groups are equally likely to see the position advertised.

In insurance pricing, we might look to offer cross subsidies to high risk demographics or locations using ideas similar to demographic parity as a justification. This practice is commonly referred to as "community rating" – equalising prices (either fully or partially) across a community irrespective of the known risk profie for any one individual. Such cross-subsidies exist within many statutory schemes in Australia, for example (though many may be better aligned to conditional demographic parity, discussed below).

Demographic parity is fundamentally different to unawareness. Unawareness only requires that our decision for any individual does not vary if we adjust the protected attribute in isolation, for that individual. It says nothing about the likelihood of such a counterfactual person

existing, or the real world treatment of protected groups *as a whole*. Under demographic parity we are requiring a form of group equality across subpopulations, but no condition is placed on the treatment of any one individual. Indeed, whilst expected outcomes may be equal for each group, under demographic parity similar individuals can be subject to quite different treatment which could be considered unfair (Dwork et al., 2012).

In order to create a situation of demographic parity (and, indeed, other group fairness criteria), we may be required to collect and utilise the protected attribute directly. Demographic parity is akin to a quota system, and quotas necessarily require knowledge of the thing that the quota is being applied to, to ensure the quota is met. In this sense, demographic parity is consistent with the concept of affirmative action, and runs counter to the notion of unawareness. We observe that quota systems are becoming increasingly common as a mechanism to create fairness, particularly in senior employment settings such as board appointments (Terjesen et al., 2014).

Demographic parity does pose another complication in that it is unclear exactly which population we ought to compute the statistics over. In a customer interaction setting, we might legitimately argue for active customers, potential customers, or some suitably weighted representation of the entire population. This problem exists for all the group fairness metrics we shall encounter in this paper, and others not covered here. Our suggestion is for practitioners to be clear on the population being used and the reasons for it, and to examine if the use of alternative populations would give materially different results.

3.4 Conditional Demographic Parity

A variant of demographic parity is *conditional* demographic parity:

$$\mathbb{E}(d \mid Z, A = a) = \mathbb{E}(d \mid Z, A = a').$$

Here we condition on a certain *legitimate* subset of factors Z in X. We might use heuristics, perhaps on the causal connection with Y, to justify which subset of attributes within X ought to be conditioned on. For example in insurance pricing we might justify using rating factors directly affecting the amount of risk taken on like sum insured, without obvious challenge. Thus we move away from the same quoted average premium across protected subpopulations if they differ in other ways that are deemed reasonable to use.

It is worth noting that once Z is granular enough such that there is no more than one member of the population for each level of Z, the definition becomes essentially equivalent to unawareness for arbitrary decision rules d.

In many situations, unawareness and demographic parity are equally appealing ideals. Conditioning on components of X to varying degrees gives us a mechanism by which we might trade these concepts off.

3.5 Calibration

Calibration (or positive predictive value parity in the binary case) requires that at each value that d could take, the expected value of the outcome should not vary by the protected attribute. Intuitively, this means any given decision "means the same thing" in terms of its relationship to Y, irrespective of A. Formally:

$$\mathbb{E}[Y \mid d, A = a] = \mathbb{E}[Y \mid d, A = a'] \quad \forall\, d \in \mathbb{R} \text{ or } \{0, 1\}.$$

We observe that this is a formalisation of the fairness definition being advocated by Northpointe, in the COMPAS debate. In the binary setting it requires that when the decision is $d = 1$ (e.g. if someone is rated "high risk" of recidivism), the odds of the true positive (e.g. actually recidivating) are equal across protected classes. We can think similarly about negative decisions.

Calibration is an intuitive concept to statisticians and actuaries. Common model diagnostics often consider plots of observed Y against various levels of d (representing either modelled or market prices). Segmenting this analysis by A (or, generally, any variable of interest) is common – this can highlight any issues of model accuracy in particular segments of the population.

3.6 Equalised Odds

Equalised odds (Hardt et al., 2016) criterion requires the decision d to be independent of A, conditional on the realised outcome Y. In the binary classification setting, we can express this as:

$$\mathbb{E}(d \,|\, A = a, Y = y) = \mathbb{E}(d \,|\, A = a', Y = y); \quad y \in \{0, 1\}$$

The moral intent of such a constraint is to ensure that A membership is not abused as a proxy for the outcome Y, whilst making sure that the "perfect" rule $d = Y$ is still admissible (sometimes said to represent the intuition that " the truth is fair").

Considering $y = 1$ alone, this requires equal true positive rates across the protected classes A. This is often termed "equal opportunity" (Hardt et al., 2016, p4). For $y = 0$, we require equal false positive rates.

This definition of fairness is appealing in some circumstances. If we are issuing bank loans, then equalised odds suggests that we should issue loans to non-defaulters with the same probability across protected classes. Similarly for defaulters, the average loan issuance rate should not vary by protected class. We note that this is a formalisation of the notion of fairness Propublica were advocating for in the COMPAS case study outlined above.

Equalised odds appears problematic in the insurance pricing setting, however, as we have observed in earlier work (see Dolman and Semenovich 2018b).

The problem can be illustrated with the following example:

- Risk type R has 10% probability of claim occurrence in the policy period, at a cost of \$1000 if a claim does occur.
- Risk type R' has the same exposure as R and an additional, independent exposure: a 1% probability of claim occurrence in the policy period at a cost of $\$10,000$. This represents the possibility of a rarer but more costly event, for example being susceptible to a natural hazard.

Let us suppose that protected types a and a' comprise R and R' in different proportions. It can then be immediately seen that under the standard model of claims process assuming independent arrivals a set of prices d fails equalised odds and similar criteria conditioning on the outcome Y, e.g for policies with no recorded claims $\mathbb{E}(d \,|\, A = a, Y = \$0) \neq \mathbb{E}(d \,|\, A = a', Y = \$0)$, unless d satisfies demographic parity: $\mathbb{E}(d \,|\, A = a) = \mathbb{E}(d \,|\, A = a')$. If the additional risk exposure represented by R' is undertaken voluntarily, demographic parity may reasonably be challenged by lower risk individuals as unfair.

Generally, where Y contains a significant element of chance, we would caution against the use of fairness criteria which condition on Y directly as chance alone ought to have no moral force.

There is another reason to avoid conditioning on the outcome that is perhaps more subtle. Let us consider again the Propublica COMPAS recidivism study, where the notion of false positive rate equality was used to make a moral statement about the fairness of the COMPAS algorithm. We ought to recognise that in this situation, a recorded observation of a follow-on offense requires a release and a subsequent new arrest. This is quite different to the underlying risk of an individual merely intending to reoffend, or actually committing another offence. It requires further that the person actually be charged and convicted. Since policing is often argued to not only be imperfect, but to also include some racial bias (Kochel et al., 2011), we should observe that not only is there some chance element, this chance is likely biased for

certain protected attributes (i.e. the chance of being caught, given you committed a crime, varies by ethnicity).

Thus the "true" intent to reoffend is reflected in the available data Y only indirectly, with considerable uncertainty and potential for large bias imposed by the policing and criminal justice process. An evaluation of a decision procedure based on such biased observations may also be biased.

It follows that a recidivism model attempting to predict the underlying intent to commit a future offence would perhaps be a better fit to the judicial problem at hand, than a model attempting to predict the available data Y. Construction of such a model would require some careful judgement and adjustments to the data Y.

Similar arguments can be made in many other situations. The available data may be only indirectly associated with the underlying moral characteristic of the individual that we intend the decision to relate to.

3.7 Actuarial Fairness

The conceptual challenges of conditioning on the outcome Y can be avoided (Dolman and Semenovich, 2018b) if we introduce the notion of (generally unobservable) *true type* μ for each individual. True type is the characteristic that has the most direct moral relevance to the decision process at hand. In the insurance pricing case, μ can represent the "pure premium" or the expected cost of claims. In the recidivism setting μ might represent the intent to commit a future crime. When we know the *type*, the morally correct decision ought to be deterministic.

There is then a mapping from μ to Y, which may also be dependent on other attributes including A. This mapping will be context dependent and may not be directly observable.

Relating premiums to the unobservable expected cost of claims has some considerable history in insurance pricing. In particular, the term "actuarial fairness" itself was perhaps first introduced by Kenneth Arrow in the essay *Uncertainty and the Welfare Economics of Medical Care* (Arrow, 1963):

> "Suppose therefore, an agency, a large insurance company plan, or the government, stands ready to offer insurance against medical costs on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean μ, the company will charge a premium μ, and agree to indemnify the individual for all medical costs. Under these circumstances, the individual will certainly prefer to take out a policy and will have a welfare gain thereby."

Arrow's work posits a certain idealised model of the market and individual decision making. Under these assumptions, full insurance coverage is the natural result. In particular, under this notion of actuarial fairness, no allowance is made for administrative expenses or profits. Thus, a more common practical interpretation of "actuarial fairness" is to describe the price, d as a function $g(\hat{\mu})$, where $\hat{\mu}$ represents the best estimate of the true risk and $\mu = \mathbb{E}(Y)$, for cost of claims Y. Typically, g is monotonic and represents loadings for expenses and profits.

We can extend this notion by definining *actuarial group fairness* as follows for expectations:

$$\mathbb{E}(d \,|\, \mu, A = a) = \mathbb{E}(d \,|\, \mu, A = a')$$

or a stronger version for distributions:

$$d \perp\!\!\!\perp A \,|\, \mu.$$

Under the first definition it is required that the expected decision is equal across protected sub-populations once we have conditioned on the type, μ. Under the second definition the requirement is for the conditional independence of d and A as illustrated in Figure 1.

Figure 1: *Left:* Graphical model representation for the equalised odds criterion in which the decision d is separated from the protected attribute A by the outcome Y. *Right:* Graphical model representation for the actuarial group fairness criterion. Now separation is through the *true type* μ.

Note that if Y corresponds to μ exactly we recover the notion of equalised odds, actuarial group fairness being a strict generalisation. The traditional concept of pure premium $d = g(\hat{\mu})$ is also consistent with this definition.

In practice, the type μ is not usually observable. Several common practices have emerged in the actuarial profession to improve the estimation of μ from observational data in the insurance setting:

1. Claim size distributions are often heavy tailed. Large claim risk may be a material contributor to μ, with the observed data Y inherently insufficient to provide an accurate estimate. To address this, a common technique is to model conditional expectation above a certain loss threshold separately with reference to industry data and expert professional judgement.
2. Claims are not always independent. Typically, some systemic risk exists across large sections of the population. For example, buildings may be subject to natural hazards which affect a large area, hence multiple policies at the same time. Historical data for such risks is naturally limited, and may misrepresent the underlying risk. Such risks, to the extent they can be identified, tend to be considered separately within the modelling process, sometimes involving industry collaboration or external expert research.
3. Claims may be subject to systemic temporal effects. For example, claim cost inflation may outstrip general inflation. Limited availability of relevant data places significant reliance on the modeller's prior beliefs and professional judgement.
4. Claims may be under-reported: they take some time to emerge and then to be settled. This makes the use of Y in an unadjusted form challenging, particularly for more recent time periods. This is of particular concern in long tailed lines. Explicit adjustments for development are often introduced.
5. Certain policy options may appear to increase the expected cost of claims while strictly reducing the policy benefits and vice versa. This occurs due to the revealed preference for the option in question itself being correlated with risk and can require adjustment to either statistical estimates or the market pricing algorithm in order that the final set of prices operate in an intuitive manner.

While much of the above is insurance specific, we suggest that this illustrates that relatively elaborate domain-specific practices may need to be developed in other settings, in order that available data Y may be used to infer the true type μ.

3.8 Causal Notions of Fairness

Notions of fairness can also be expressed in causal language. Several formalisations of causality exist (e.g. Imbens and Rubin, 2015; Pearl, 2009) which we do not attempt to discuss here.

One approach to causality is through counterfactual statements. Denote $d_{A \leftarrow a}$ the outcome of a decision d if the value the protected attribute were set to the value of a. While much of the literature leaves the nature of such hypothetical interventions unclear (with notable exceptions, e.g. Greiner and Rubin, 2011), it suffices to note that the values of X that causally depend on A are also assumed to change to x'.

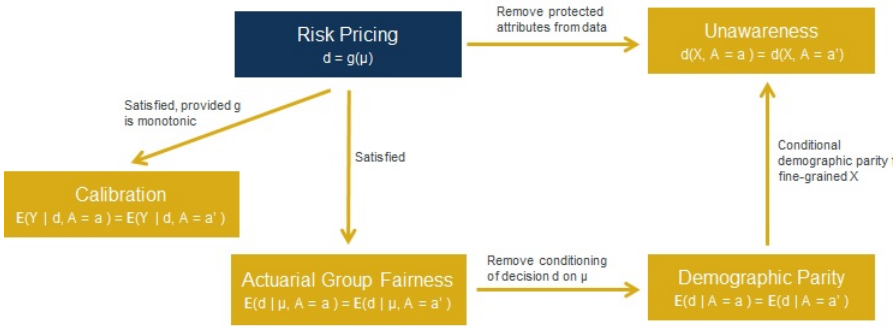

Figure 2: *Some relationships between fairness concepts*. There are many connections between the concepts discussed, enabling us to "move" (at least definitionally) from one fairness criterion to another. Construction of procedures to modify a decisioning algorithm to enact such transitions are beyond the scope of this paper, but would be a useful area of future research.

Individual counterfactual fairness can be definied as

$$d_{A\leftarrow a}(x, a) = d_{A\leftarrow a'}(x', a')$$

and requires that for every individual the decision is the same regardless of the hypothetical intervention on A.

Conditional counterfactual fairness is a weaker condition in that

$$\mathbb{E}(d_{A\leftarrow a}(x, a) \mid Z, A = a) = \mathbb{E}(d_{A\leftarrow a'}(x', a') \mid Z, A = a)$$

for a subset of factors Z in X. It requires that the decisions under the hypothetical intervention on A are equal in expectation for strata defined by Z. If Z is sufficiently granular we end up with individual fairness, just as in the earlier example with conditional demographic parity and the unawareness criteria.

An insurance example where causal reasoning is illuminating is shown in Kusner et al. (2017). They consider a motor rating scenario where the choice of car color X is caused by an unobserved factor U (e.g. "aggression") which is causally linked to the risk of accidents. There is a protected factor A causally linked to the choice of color but which has no effect on risk. If we represent causal links as arrows, this yields $A \rightarrow X \leftarrow U \rightarrow Y$. In this scenario, using only X for rating (analogous to unawareness) may be viewed as unfair as it will disproportionately impact certain strata of A which is disallowed under conditional counterfactual fairness.

3.9 Metric Incompatibility

Following the Propublica article, various researchers explored the compatibility of fairness metrics. Almost at the same time, Kleinberg et al. (2017) and Chouldechova (2017) found that in situations like the Propublica example (i.e. with different base rates of Y across protected populations), only two out of three of the following could be achieved:

- Calibration
- Equalised odds for $Y = 0$
- Equalised odds for $Y = 1$

Research has further suggested that this issue of incompatibility extends to other fairness metrics that can be derived from common observational statistics. It has generally been accepted that trade-offs are essential and we cannot hope to address all considerations simultaneously.

This problem of incompatibility has important implications for any regulatory or customer advocacy body seeking to create a generalised situation of "fairness" across diverse decision settings. Such an intent, whilst reasonable as a high level aim, is clearly not specific enough to guide the design of a decisioning algorithm. To give greater specificity requires in-depth,

expert analysis of each decision function, the data available to produce it and the desires and expectations of parties involved. We suggest this is not a scenario that is amenable to simple rules.

Understanding the relationships between different fairness metrics could be the first step towards the assessment of possible trade-offs. Some of these relationships are illustrated in Figure 2.

4 Analysis of Insurance Pricing Practices - Links to the Fairness Research

In the previous section we have examined some fairness concepts from the recent literature. Now we attempt to investigate several common insurance pricing practices and methodologies in light of this research.

4.1 Variation from Pure Risk Estimates

The traditional notion of actuarially fair price is to set the premium d equal to the expected cost of claims, μ, perhaps allowing for expense and profit loadings. We know, however, that μ cannot be observed directly. Indeed, it is often the role of actuaries to provide an estimate $\hat{\mu}$ for μ.

Should the price d always be set exactly equal to $\hat{\mu}$? Common practice is to admit some variation.

This can in part be justified by the inherent uncertainty of $\hat{\mu}$. In particular, if $\hat{\mu}$ is unreliable, then a price equal to it will likely be unstable, especially if $\hat{\mu}$ is re-estimated frequently. This in itself may have an adverse impact on policyholders and some limitation on this instability may prove desirable.

A natural question to ask is how much variation should be admitted as a response to this instability? One approach is to use credibility to blend the estimates. Assume that there are two periods with $\hat{\mu}_0 = \mu + \epsilon_0$ and $\hat{\mu}_1 = \mu + \epsilon_1$, with

$$\begin{aligned} \epsilon_0 &\sim \mathcal{N}(0, \sigma_0), \\ \epsilon_1 &\sim \mathcal{N}(0, \sigma_1). \end{aligned}$$

Then the optimal $\hat{\mu}$ at time 1 is given by the standard credibility formula $(1-w)\hat{\mu}_0 + w\hat{\mu}_1$ with

$$w = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2}}.$$

This can be interpreted as adjusting the old estimate towards the new one, with the size of adjustment dependent on the relative reliability of the two estimates. In practice this is often substituted for a simple mechanism limiting the absolute value of allowed percentage premium change year on year. This naturally creates variation away from the idealised $d = \mu$.

We must note that the variation in the estimate $\hat{\mu}$ is not always driven by an independent noise process – systematic externalities do emerge. Superimposed inflation is a common example – emerging risk of superimposed inflation might justify a significant departure from $\hat{\mu}$ to allow for anticipated adverse development. Incorporation of new sources of information is another common source of variation between estimates.

4.2 Group Fairness

Examination of one-way charts (Figure 3) is a common element in both the construction of claims cost models and portfolio performance reviews.

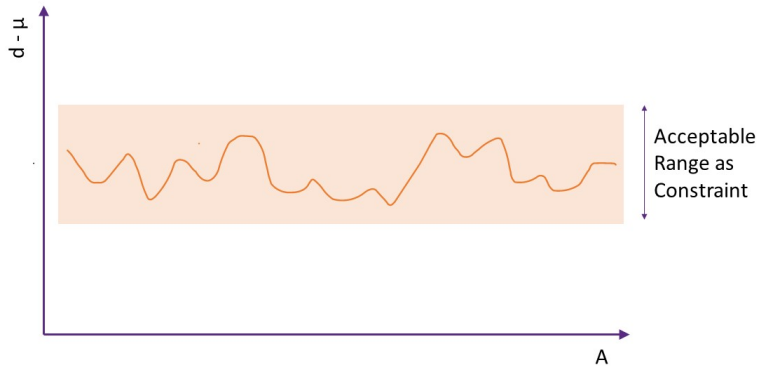

Figure 3: If we examine the difference between the market price d and the cost estimate $\hat{\mu}$, we can impose a constraint on the range of the outcome, as depicted in this stylised example.

Typically, values for the observed claims Y, the risk estimate $\hat{\mu}$ and the market price d may be plotted against each rating factor, for a suitable reference population (e.g historic business written) to check for unusual trends or divergence across the levels of the rating factor.

We might adapt this common practice into a formal constraint over a pricing algorithm, by using the notion of *actuarial fairness* we introduced earlier. The existence of allowable variation in the decision d due to noisy estimates of μ may necessitate a weakening of equality, perhaps to that of an allowable range. The use of ranges is also likely to be more practical in situations where the firm is not in full control of the outcome (for example if group fairness notions are evaluated using quotes offered, the composition of the evaluation group is unknown at the point of price setting).

4.3 Restricted Factors and the Use of Offsets

A common practical approach for dealing with "protected" variables, (e.g. gender under the EU Gender Directive), is to remove the variable in question from the data and update the model, allowing partial compensation via correlated factors.

A related methodology is the use of offsets within the generalised linear modelling framework to constrain the parameters for an explanatory variable to take certain values. This is commonly applied for items like bonus/malus schemes, where relativities are published in advance and are not usually open to change, forcing the corresponding regression parameters to take published values while parameters for correlated factors are varied to partially compensate. For further background reading including technical derivations, see Anderson et al. (2007) and Semenovich (2013).

This approach to restricted factors is consistent with the "unawareness" criterion described above, where we insisted that a change in protected attribute status for any one individual had no impact on the outcome. Use of offsets to constrain coefficients to certain values can be viewed as a generalisation, with "unawareness" a special case with these values set to zero.

The analogy is important to recognise. As mentioned earlier, unawareness is vulnerable to proxies for the protected attribute. Offsets are equally vulnerable – indeed, when used directly, the offset technique will force any correlated parameters to adjust in an attempt to remove any bias imposed by the offset.

This may or may not be an issue depending on the intended use of the model. Werner and Guven (2007) describe this issue and an alternative approach – allowing the model to be free-fitted and manually adjusting relativities outside of the modelling process. We note that this is consistent with counterfactual fairness criterion introduced by Kusner et al. (2017) under

the linearity assumptions introduced in their paper and can also yield models satisfying conditional demographic parity.

5 Conclusions

Practical Suggestions

In this paper, we have reviewed several fairness criteria that are regularly seen in the contemporary fairness literature, outlining some of the benefits and drawbacks of their application. We note that many of them already have some involvement in insurance markets, albeit usually with different names and historical justification for their use.

A critical observation is that it is generally not desirable to be favourable towards one of these criteria in all circumstances - each has clear flaws in certain situations. Additionally, it is not usually obvious which metric should be preferred in any given situation – good minds will often disagree about the correct outcome.

Furthermore, many of these metrics are incompatible. We should observe that there are many more potential definitions of fairness other than those discussed here, and some definitions of fairness that use entirely different logic. The problem of incompatibility generally grows as we introduce more potential ways to define fairness. This is likely to create challenges for any future legislation in this area. It certainly creates challenges for practitioners who must determine what to do, today - we cannot meet all ideals simultaneously.

In these circumstances, we suggest the following steps for actuaries and data scientists to consider when developing decision making algorithms, in order to incorporate fairness as a consideration:

1. **Create Clarity**

   Firstly, when deploying a decision making algorithm we should be clear on why we consider the decision procedure to be fair and reasonable, and formalise this somehow in our businesses. This should include considerations of group fairness such as those outlined above, as well as considerations of individual fairness. If we have already considered fairness in this way, we should ensure it is clearly documented and all relevant staff are aware of the considerations and the decision. If not, we ought to consider instigating such a discussion in our businesses.

2. **Acknowledge Imperfections**

   Secondly, we should acknowledge the inherent tradeoffs required. As noted above, due to incompatibility between metrics it is perhaps not sufficient to operate with a high level commitment merely to "act fairly and reasonably". Whilst this might appear fine as a high level principle, for algorithmic decisions this is clearly insufficient. We need to be far more precise about how we apply concepts of fairness to each situation, and acknowledge the inherent trade-offs being made.

3. **Be Adaptable**

   Thirdly, it would be unreasonable for us to merely state that after some debate, we will always define fairness in a particular way across our enterprise. If nothing else, this runs the risk that a superior notion of fairness is developed following such a decision. The justification ought to be specific to the decision at hand, and ought not to be held too strongly: we will likely need to adapt any answer over time, particularly as the research environment matures.

4. **Act With Humility**

   Finally, we should acknowledge that people can genuinely hold differing views on the "right answer" for questions such as these. Hence, whatever we do, we may be called to change our approach, and if and when this occurs, we ought to be open to the discussion.

A commitment to openly discuss views which may contradict our own, and a commitment to rectify any issues as they are identified, and adapt according to society's evolving norms, appears the only reasonable course of action.

Opportunities for Actuaries

It is increasingly clear that effective application of ideas of algorithmic fairness requires significant professional judgment and domain knowledge, defying any simple prescriptions.

In general, the recent emergence of algorithmic fairness as both an academic discipline and an area of public policy discussion should be exciting for actuaries. As a profession, we have skills, training and a long standing tradition of carefully considering fairness, in our (somewhat narrow) domains of expertise, where algorithmic decisions have been present for some time. As algorithmic decisions are increasingly applied in other domains, we should see this as a great opportunity for actuaries to broaden their remit, showing the value of professional judgement and ethics in quantitative disciplines.

We hope that by demostrating links between recent academic theory and common insurance practice familiar to many of us, more members of the profession will feel comfortable entering into the broader fairness debate and into analytics in wider fields.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L., Machine bias, ProPublica (2016)

Arrow, K., Uncertainty and the Welfare Economics of Medical Care, American Economic Review, Vol. 53, No. 5. (1963)

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N, A Practitioner's Guide to Generalized Linear Models, A CAS study note, Third Edition, February 2007, Casualty Actuarial Society (2007)

Brennan, T., Dieterich, W. and Ehret, B., Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System., Criminal Justice and Behavior 36: 21 (2009)

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

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S., A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear, Washington Post (2016)

Dieterich, W., Mendoza, C. and Brennan, T., COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Northpoint Inc. Research Department (2016)

Dolman, C., Semenovich, D., Algorithmic Fairness: Contemporary Ideas in the Insurance Context, Institute and Faculty of Actuaries GIRO Conference (2018a)

Dolman, C., Semenovich, D. Actuarial Fairness. Workshop on Ethical, Social and Governance Issues in AI, NeurIPS 2018 (2018b)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., Fairness through awareness., In Proc. ACM ITCS, pages 214–226 (2012)

Greiner, D.J. and Rubin, D.B., Causal effects of perceived immutable characteristics., Review of Economics and Statistics, 93(3), pp.775-785. (2011)

Hardt, M., Price, E., Srebro, N. et al., Equality of opportunity in supervised learning, Advances in Neural Information Processing Systems (2016)

Imbens, G., and Rubin, D., Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press (2005)

Kleinberg, J., Mullainathan, S. and Raghavan, M., Inherent trade-offs in the fair determination of risk scores, Proceedings of Innovations in Theoretical Computer Science (ITCS) (2017)

Kochel, T., Wilson, D. and Mastrofski, S. , Effect of suspect race on officers' arrest decisions, Criminology: An Interdisciplinary Journal, 49(2), 473-512 (2011)

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., Counterfactual fairness, In Advances in Neural Information Processing Systems (2017) (pp. 4066-4076)

Liu, L., Dean, S., Rolf, E., Simchowitz, M. and Hardt, M., Delayed impact of fair machine learning , International Conference on Machine Learning (2018)

Pearl, J., Causality., Cambridge university press (2009).

Pedreshi, D., Ruggieri, S. and Turini, F., Discrimination-aware data mining, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (2008)

State of Wisconsin vs Loomis, Supreme Court of Wisconsin (2016)

Terjesen, S., Aguilera, R., Lorenz, R., Legislating a woman's seat on the board: Institutional factors driving gender quotas for boards of directors, Journal of Business Ethics (2014)

Werner, G. and Guven, S., GLM Basic Modeling: Avoiding Common Pitfalls, Casualty Actuarial Society Forum, Winter 2007