Algorithmic Fairness: Contemporary Ideas in the Insurance Context

C. Dolman, D. Semenovich

Abstract: We review contemporary ideas of quantitative algorithmic fairness as explored by the computer science academic community, to improve member awareness of this recent research. We identify relationships between these ideas and more traditional actuarial concepts, notably from insurance pricing and provide high level guidance for practitioners on how to adopt some of the concepts here in their work.

1 Introduction and Motivation

As societies have 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 knowledge of it should still prove valuable to actuaries working with algorithms, particularly as broader study directed towards the topic may require us to re-examine our historic normative practices. Improving awareness of this research is our primary objective.

There is also an opportunity for actuaries to contribute to the emerging societal debate. Insurers have long-standing norms around the fairness of insurance pricing, which is clearly a special case of an algorithm, and the profession's deep consideration of this topic over many years may be of substantial value, including outside of insurance. We aim to show links between the recent academic literature and common insurance practices in order that actuaries might better participate in the current debate. In doing so, we have identified some extensions of the contemporary research which feel might warrant some further attention, either from the research community and actuarial profession.

2 Fairness in Legislation

We begin by briefly exploring existing concepts of fairness, particularly those encoded within anti-discrimination acts. We do not profess to be legal scholars so our discussion here should be taken as the high level views and interpretations of lay practitioners.

Many countries have enacted anti-discrimination laws, protecting people from being disadvantaged based on certain characteristics. Common characteristics that are legislated in this manner include age, gender, sexuality, disability, religious affiliation and political affiliation. In this paper we shall generally refer to such characteristics as *protected attributes*, in line with the common vernacular of the research community.

One common way to categorise discrimination norms and laws is by differentiating between direct and indirect discrimination:

1. **Direct Discrimination** – an act of direct discrimination requires us to actively use the protected attribute to cause disadvantage to an individual, relative to another. An example would be a company refusing to consider female applicants for a job advertisement, just for being female.

2. **Indirect Discrimination** – an act of indirect discrimination requires us to take an action which gives rise to disadvantage to a protected class, but without the direct use of the protected attribute. An example would be a company adopting a minimum height of 1.8 metres for future employees when advertising a role, where height is not relevant to success in the role. This would naturally give rise to fewer females being eligible than males.

   Indirect discrimination can sometimes be allowed under legislation, particularly if avoiding indirect discrimination would be overly burdensome. As might be expected, the boundaries of this allowance have been the subject of great debate.

We observe that these two concepts are not always compatible. The clearest example is that of "affirmative action" – the deliberate attempt to counter potential indirect discrimination (or other societal bias with similar effect) via the direct use of the protected attribute as a counterweight in decision making. By definition, therefore, affirmative action utilises a protected attribute for decision making, contrary to the ideal of avoiding direct discrimination. A recent high profile example is the use of gender quotas for corporate board seats – in order to counteract the predominance of male appointees, quotas for females have been adopted in many organisations and have been legislated in some countries (Terjesen et al., 2014). Affirmative action is sometimes also referred to as "reverse discrimination", and is regularly challenged in courts: for discussion of a recent example in the education sector, see Benner (2018).

The notion of affirmative action, and the general issue of compatibility of these two concepts of discrimination, has some relevance to algorithmic decision making. In particular, many recent notions of algorithmic fairness are akin to notions of indirect discrimination, and many proposed approaches to ensure compliance with them require direct use of the protected attribute (for example, the post-processing step introduced by Hardt et al., (2016). Thus, modern notions of algorithmic fairness are potentially open to criticism on similar grounds to affirmative action.

We also observe that many laws were written in an era prior to the mass use of algorithmic decision making. In general, a clear act by an individual human being (or human led enterprise) could be identified and debated as to its compliance with the law. As has been highlighted by several academics recently, algorithmic discrimination creates a number of challenges to the structure of current laws. Much of the commentary and research has been in relation to US law, though the themes and issues raised may generalise to other countries. For a good introduction, see Barocas and Selbst (2016).

It is finally worth noting that many existing anti-discrimination acts contain special exemptions for insurance contracts (Swiss Re, 2011). In many cases the insurance industry was able to successfully argue that moving away from risk pricing would increase the risk of adverse selection and consequent market failure, thus justifying the use of risk-predictive attributes such as gender and age which would otherwise be protected by the proposed anti-discrimination legislation.

This stance has remained stable for some time, however it is possible that if new, legislation

around algorithmic decision making should be introduced, insurance may not qualify for the same range of exemptions as might have applied previously. Changing standards are already evident, for example in the 2011 European Court of Justice ruling regarding the validity of insurance derogation for the "unisex rule" of the 2004 EU Gender Directive (Rebert & Van Hoyweghen, 2015).

3 The COMPAS Debate

The recent academic research was motivated by several high profile examples. We will examine one of them in some detail to illustrate relevant considerations.

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 called 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 probation 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 allegation based on data Propublica had obtained from Broward County, Florida. Propublica observed 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 allegation 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 "higher 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. 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 very narrow, trivial, circumstances (Kleinberg et al. 2017, Chouldechova 2017). Tradeoffs are necessary.
- 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. However, the adversarial nature of the public discussion was perhaps not conducive to a debate over trade-offs that the situation demanded. COMPAS is still being used today.

4 Formal Definitions of "Algorithmic Fairness"

Cases such as COMPAS illustrate that even though there are societal norms, and indeed laws, against discrimination, by constructing the decision in a mathematical form we demonstrate that we may need to think more precisely about exactly what norms we are looking to encode and enforce.

Following this, many different notions of algorithmic fairness have been proposed by the research community. In this section, we review several that appear of immediate relevance to the insurance industry and translate them into insurance terms. We also discuss the traditional notion of "actuarial fairness" and how it can be generalised to established connections with other proposals in the recent literature.

4.1 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 (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.

4.2 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:

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

In some sense this parallels the concept of "direct discrimination" which is commonly legislated. 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, we do not have access to a protected attribute. Hence we are often under a situation of "unawareness" by default.

Owing to the 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 (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 the various notions of indirect discrimination that we shall come to below. As noted in Section 2, most proposals for correction mechanisms to ensure algorithmic fairness use the protected attribute directly as part of a corrective process.

4.3 Demographic Parity

A more demanding fairness metric is "demographic parity" or "statistical parity". This concept requires the probability of a positive decision to be equal across protected classes.

We can express this idea formally as:

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

noting that for binary classification $\mathbb{E}(d) = \Pr(d = 1)$. At one level this is appealing. If d is the decision of a bank to lend money, for example, this metric requires the probability that a loan is given to be the same, irrespective of the membership of a protected class a. However, this particular case might lead to complications, for example if a particular protected group is generally not creditworthy and hence more prone to default. In this situation, the use of a decision procedure complying with demographic parity might lead to greater harm

to the protected group than other decision procedures. 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 ad.

In insurance pricing, we might look to offer cross subsidies to high risk demographics or locations using ideas similar to demographic parity (or some simple derivation from it) as a justification. Such cross subsidies exist within many statutory schemes in Australia, for example (though some may be more aligned to conditional demographic parity, discussed below).

Demographic parity is fundamentally different to unawareness. Unawareness just requires that our decision for any individual does not vary if we adjust the protected attribute. It says nothing about the treatment of each group *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, individuals can be subject to quite different treatment which could be considered unfair (Dwork et al., 2012).

Furthermore, to create a situation of demographic parity (and, indeed, other group fairness criteria), we may be required to be aware of 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. In this sense, demographic parity is consistent with the concept of affirmative action, and runs counter to the notion of unawareness.

Demographic parity does pose a slight complication, in that it is unclear exactly which population we ought to compute the statistics over. We might legitimately argue for active customers, potential customers, or some representation of the entire population. This problem exists for all the group fairness metrics we shall encounter in this paper. Our suggestion is for practitioners to be clear on the population being used and the reasons for it.

4.4 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 attributes within X ought to be conditioned on. For example in insurance pricing we might justify using rating factors 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 *legitimate* to use, under some justified definition of *legitimate*.

It is worth noting that this definition becomes essentially equivalent to unawareness if Z is sufficiently rich. Thus, we have defined a sliding scale from unawareness at one end, to full demographic parity at the other, with the level of conditioning on subsets of X defining the position on the scale. Once Z is granular enough such that there is no more than one member of the population for each level of Z, it is 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 vehicle in which we might

trade these concepts off.

4.5 Equalised Odds

Equalised odds (Hardt et al., 2016) is another intuitive concept when considering specific examples of its implementation. It requires the decision d to be independent of A, conditional on the realised outcome Y. If we use the simple binary classification setting, this is formalised as:

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

Considering $y = 1$ alone, this constraint requires equal true positive rates across the different demographic groups. For $y = 0$, we are requiring equal false positive rates. The equalised odds criterion requires both true positive and false positive balance to hold at the same time. We can also consider them as separate criteria in isolation.

This definition of fairness is intuitively 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, and similarly for defaulters. We note that this is a formalisation of the sort of fairness Propublica were advocating for, in the COMPAS debate outlined above.

The equalised odds definition is, however, harder to argue intuitively within insurance pricing than unawareness or demographic parity. Conditioning directly on Y appears unnatural as there is a significant component of chance in individual claims outcomes. For example, it is not self-evidently fair to require average premiums to be the same for policies with no claims (observed ex-post) across protected subpopulations. Under the standard model of claims process with independent arrivals this is essentially equivalent to demographic parity.

4.6 Actuarial Fairness

The concept of "actuarial fairness" in insurance pricing goes back a long way. Martinéz et al. (2016) identify scholarship in the area of fair pricing going back at least to the 17$^{\text{th}}$ century, when the principle of equality in risk emerged in the modern tradition. For example, Jan De Witt (Haberman & Sibbet, 1995) argued in 1671 that the relative price of lifetime annuities ought to reference the relative chance of death, thus following contemporaneous work on valuing risky gambles: a fair price for uncertainty is the expected value of the outcome.

In the subsequent era, this concept was referred to by a variety of terms, such as "pure premiums", "just premiums" and "equitable rates" in the actuarial literature. Of note was the lengthy debate around the use of so called gross premium or net premium valuation approaches for life insurance contracts, primarily centring around how expenses, bonuses and profits, quantities not readily attributable to individual policies, ought to be equitably recognised and distributed as experience emerges (Turnbull, 2017).

Significant progress was made in the second half of the 20$^{\text{th}}$ century applying tools of neoclassical microeconomics to the study of insurance markets. In particular, the term "actuarial fairness" itself was perhaps first introduced by Kenneth Arrow in his acclaimed 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.

We propose a new criterion which we term *actuarial group fairness*, that both generalises equalised odds for binary classification and is appropriate for use in situations involving a degree of randomness, like insurance pricing. We do this by introducing the concept of unobserved *type* μ associated with each member of the population.

Type can correspond to the intent to reoffend, or to commit fraud, or in the case of insurance pricing can represent the "true" expected cost of risk (as described above). *Type* should represent the attribute that is of direct moral relevance to the decision at hand. The realised outcome Y can then be interpreted as a noisy (and potentially biased) realisation of the process that μ intends to describe.

We formally define *actuarial group fairness* as:

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

for all values of μ.

Under this definition, we require that the expected decision is equal across protected sub-populations once we've conditioned on our notion of individual *types*.

If Y corresponds to μ exactly, we recover the original definition of equalised odds. We claim that this would be true in situations where an individual has perfect agency over the realised outcome Y: their type becomes realised with perfect accuracy. Thus, our definition of Actuarial Group Fairness is a strict generalisation of the concept of equalised odds, to include settings where Y might not fairly reflect the underlying *type* of an individual.

In many situations, we would suggest that an individual does not have perfect agency over Y. In insurance pricing this is self-evident: chance plays a significant role in the claims process.

In the criminal justice setting we used earlier as a motivational example, we observe that Y represents the individual being found to have reoffended by the authorities. This is quite different to the underlying risk of an individual actually intending to reoffend: it requires that the person be charged and convicted of an offence. Since policing is often argued to include some racial bias (Kochel et al., 2011), we should observe that not only is there some chance element in the policing process i.e. $\mathbb{E}(Y \mid \mu = 1) < 1$, but that this chance is likely biased for certain protected attributes i.e. $\mathbb{E}(Y \mid \mu = 1, A = a) \neq \mathbb{E}(Y \mid \mu = 1, A = a')$.

It follows that a recidivism model attempting to predict the underlying intent to commit an offence (which we might deem the *type*, or μ), would perhaps be a better fit to the judicial problem at hand, than a model attempting to predict Y. Construction of such a model would require some careful judgement, since μ is unobserved, but perhaps might be inferred from some duly adjusted Y.

We note that an argument of the form above, surprisingly, appears to be missing from

the current literature (which seems to implicitly assume Y is reflective of what we have deemed the underlying *type*). This idea perhaps warrants further exploration and formalisation.

In the case of insurance pricing, *actuarial group fairness* requires that the market premium is expected to be the same for policies with the the same "true" risk cost regardless of membership of protected subpopulations. Hence risk pricing strategies, or simple derivations from them, are likely to be sufficient to meet the definition. This includes traditional insurance notions of "actuarial fairness" like those of Arrow noted above, which become special cases that are sufficient to comply with our more general notion. Formally, setting $d = g(\mu)$ for some function g is sufficient to satisfy our definition, though other solutions are also viable.

In practice, the "true" risk cost is not directly observable, but would normally be estimated by a technical pricing model that may include significant professional judgement (i.e. we propose that in practice one would condition on the best estimate $\hat{\mu}$ rather than μ). So, consideration of a fair construction of $\hat{\mu}$ is also warranted: again, we suggest this could be a useful extension to our work.

4.7 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:

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

for all d.

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

In the case of insurance pricing, this translates into the average actual cost per policy being equal across protected groups for each premium level in market. If we adopt risk pricing as before i.e. setting $d = g(\mu)$, and further insist that g is monotonic, this is sufficient to (asymptotically) comply with this criterion. For covers resulting in low frequency but high severity claims, calibration may be impossible to validate empirically, even as an approximation.

4.8 Calibration and Equalised Odds: Incompatibility

Following the Propublica debate, 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 $Y = 1$

Indeed, research has further suggested that this issue of incompatibility extends to other notions of fairness that can be derived from common observational statistics. Thus, it has generally been accepted that in defining and applying ideas of fairness, we will have to make trade-offs and cannot hope to satisfy all potential ideals simultaneously.

4.9 Weaker Notions of Actuarial Group Fairness and Calibration

In most case studies from the literature, notions of fairness such as those discussed above are used in an aspirational manner. It is almost never the case than an algorithm meets them strictly: some approximation to equality is loosely deemed to mean an algorithm is compatible with the fairness criterion.

In the insurance pricing setting we are dealing with continuous values of μ and d and the notions of fairness which insist on precise equality are not immediately applicable. This can be addressed by substituting weaker approximate criteria instead.

Let K and δ be some values in $\mathbb{R}^+$. Then we define *weak actuarial group fairness* as:

$$\left\| \mathbb{E}\Big[d \,\Big|\, K(n-1) \leq \mu < Kn, A = a\Big] - \mathbb{E}\Big[d \,\Big|\, K(n-1) \leq \mu < Kn, A = a'\Big] \right\| < \delta$$

and *weak calibration* as:

$$\left\| \mathbb{E}\Big[Y \,\Big|\, K(n-1) \leq d < Kn\,, A = a\Big] - \mathbb{E}\Big[Y \,\Big|\, K(n-1) \leq d < Kn, A = a'\Big] \right\| < \delta$$

for all $n \in \mathbb{N}$.

These are conceptually simple: rather than seeking equality in expectation for members of the population that we define as identical, we weaken this to a notion of similarity (an absolute difference of δ) for a group of *similar* members of the population, with respect to a partition of $\mathbb{R}$ into K-sized slices.

Since we have control over the decision d, it is simpler to maintain compliance with the notion of weak actuarial fairness, than the notion of weak calibration.

4.10 Relationships Between Concepts Discussed

The schematic in Figure 1 summarises the concepts discussed above and their interrelationships identified.

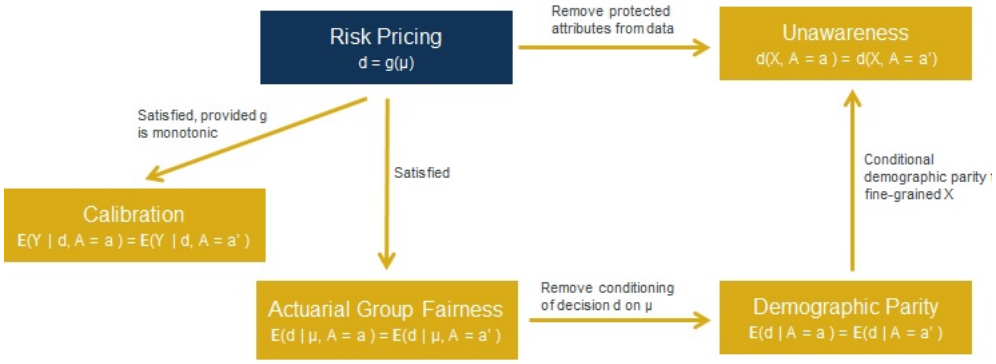

Figure 1: Some relationships between different concepts of fairness.

There are many relationships between all the concepts discussed above, 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.

5 Conclusions: What Should Actuaries Do?

Recent discussions of algorithmic fairness do have many parallels in the norms of the insurance industry when it comes to pricing:

- "Unawareness" of protected attributes is common, either due to it being mandated by law or because the data is simply not collected in any case (giving us unawareness by default).
- Regimes that mandate full or partial "community rated" equalisation of pricing (for example some statutory schemes in Australia) appear consistent with ideas of (conditional) demographic parity, at least to some degree.
- Strict "risk pricing", sometimes termed "actuarial fairness", is a special case of what we have defined here as "actuarial group fairness", which is a generalisation of the notion of equalised odds.
- Strict risk pricing also appears compatible with notions of "calibration".
- Many pricing systems in market are likely to be hybrids of these and other ideas, and may well translate into hybrid/tradeoff states in the modern formalism. Understanding this and how to formalise the many options available to an insurer could be a useful extension of our work.

Whilst many actuaries tend to focus on pricing above other problems, some are beginning to turn their attention to the analysis of other decisions, making the academic discussions around fairness in the binary classification setting directly applicable.

Some examples of binary decisions that actuaries may work on include:

- Whether to send a cross sell marketing offer to an individual, or not.
- Whether to offer a discretionary discount or option at sale time, or not.
- Whether to offer an ex-gratia payment or other offer (for example offering a hire car where there is no policy option) for a claim, or not.
- Whether a claim should be referred to a fraud investigations team, or not.

We think it is reasonably likely that most forms of binary decision undertaken by insurers are being investigated for algorithmic automation by someone, somewhere, and that actuaries may well be involved in this process. Fairness should be actively considered as part of such projects. So what should we do, as a profession, in this rapidly evolving environment? In our view it would be prudent to act in four related ways:

1. **Create Internal Clarity**

   Firstly, we should be clear on why we consider our actions 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. It is not possible to be all things to all people, meeting all potential definitions of fairness simultaneously. Indeed, in some circumstances, this has already been demonstrated mathematically. A high level commitment merely to "act fairly and reasonably" might appear fine as a whole, but for algorithmic decisions this is clearly deficient. 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, one size will not necessarily fit all. 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.

We encourage members of the profession to contribute to the emerging research environment and public debate. The actuarial profession has a long tradition of research into fairness in our domains of expertise, and we believe this tradition can and should be harnessed for broader public benefit.

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)

Barocas, S. and Selbst, A., Big Data's Disparate Impact, 104 California Law Review 671 (2016)

Benner, K., Justice Dept Backs Suit Accusing Harvard of Discriminating Against Asian-American Applicants, New York Times (2018)

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)

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

Haberman, S. and Sibbet, T. A., The history of actuarial science, William Pickering, London (1995)

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

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)

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

Martinéz, A. J. H., Teira, D. and Pradier, P., What was fair in actuarial fairness. Documents de travail du centre d'e´conomie de la Sorbonne (2016)

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)

Rebert, L. and Van Hoyweghen, I., The right to underwrite gender: The goods & services directive and the politics of insurance pricing, Tijdschrift Voor Genderstudies, 18, pp. 413–431. (2015)

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

Swiss Re., Fair Risk Assessment in Health & Life Insurance, Swiss Re, Zurich (2011)

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)

Turnbull, C., A History of British Actuarial Thought, Palgrave Macmillan (2017)