



Actuaries Summit

Time to Act

3–4 June 2019



**Actuaries
Institute**

International Convention Centre Sydney

Algorithmic Fairness: Some Practical Considerations for Actuaries

Chris Dolman, Dimitri Semenovich

© Chris Dolman, Dimitri Semenovich

*This presentation has been prepared for the Actuaries Institute 2019 Actuaries Summit.
The Institute Council wishes it to be understood that opinions put forward herein are not necessarily those
of the Institute and the Council is not responsible for those opinions.*

Background and Aims

- Extension of our previous work from GIRO and NeurlPS last year
- Aim is threefold:
 - Improve awareness of recent academic research into “algorithmic fairness”
 - Draw links between this research and common insurance industry practices
 - Illustrate some opportunities for actuaries, and give high level guidance for considering fairness in algorithms
- More in the paper than we can cover today – please read it!

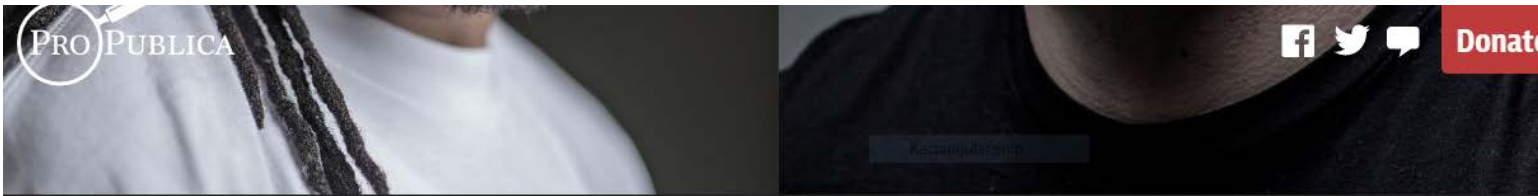
Motivational Example - COMPAS

- Many algorithms used in criminal justice settings around the world
- COMPAS is the most (in)famous one – there are many others
 - **C**orrectional
 - **O**ffender
 - **M**anagement
 - **P**rofilng for
 - **A**lternative
 - **S**anctions
- Aim is to provide judges with an objective, data driven “risk score” for different forms of recidivism, to help inform decisions
- Final decision for things like bail, parole, etc still sits with the judge.

Scandal?!





Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

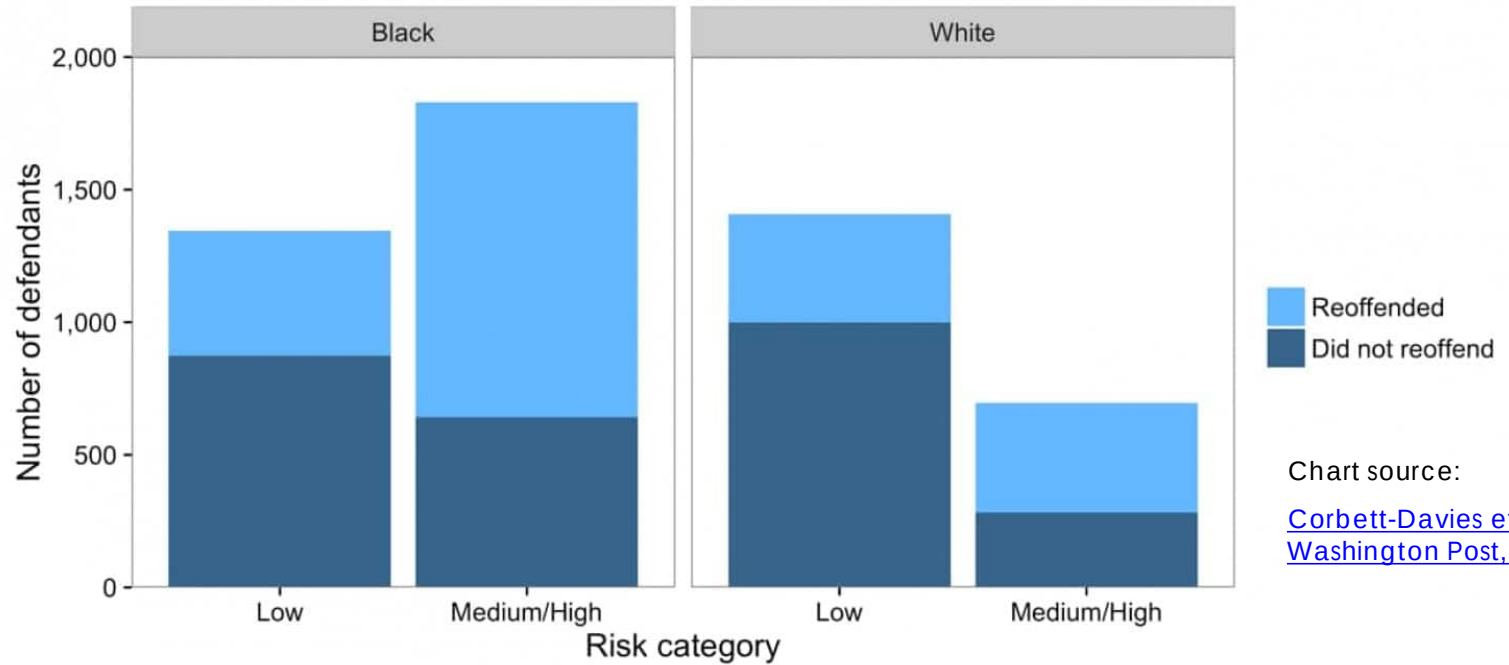
Claims Made and the Response

Propublica obtained the COMPAS risk scores of people in Broward County, Florida.

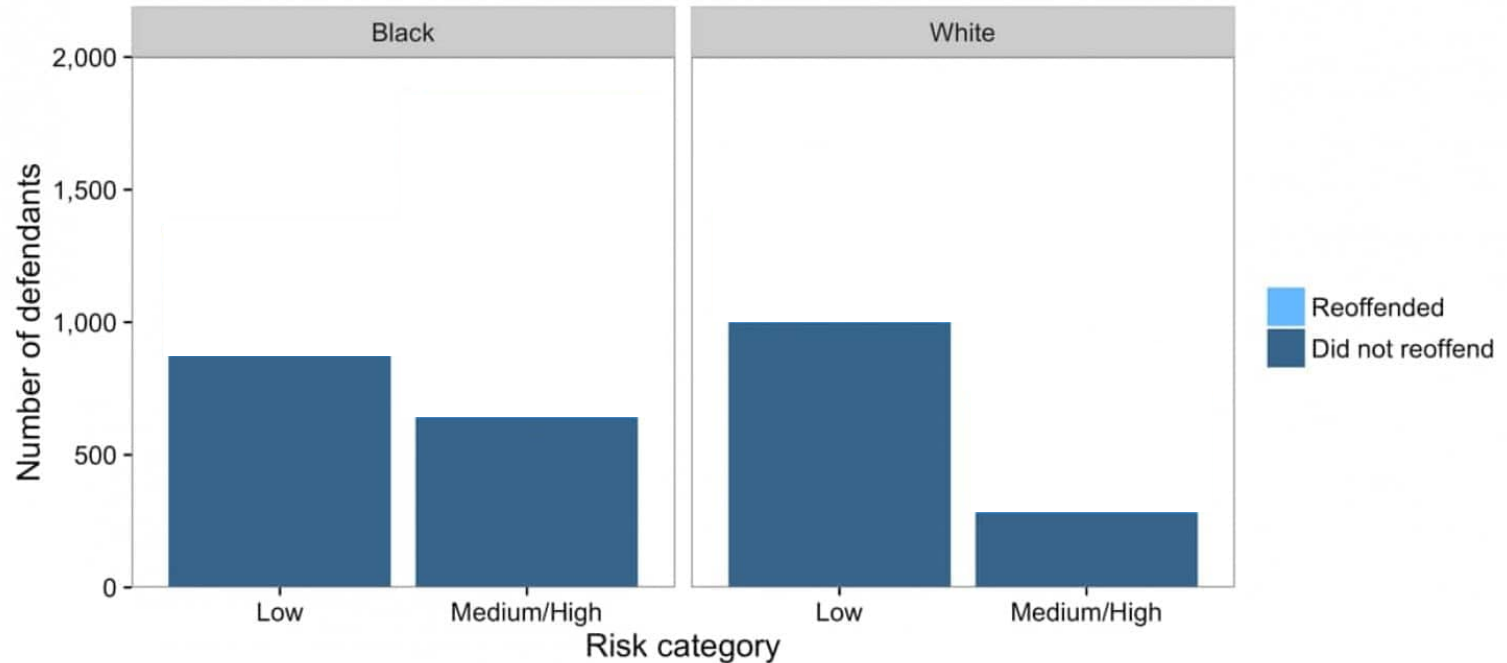
- Grouped scores into “low” and “med/high” risk
- Checked how many people were actually charged with new crimes in the two years after the risk scoring, compared the results to the risk scores
- Claimed: “Black defendants were often predicted to be at a higher risk of recidivism than they actually were” ([Larson et al., 2016](#))

Makers of COMPAS (Northpointe Inc.) claimed the algorithm was in fact fair, using a different definition of fairness

The Data

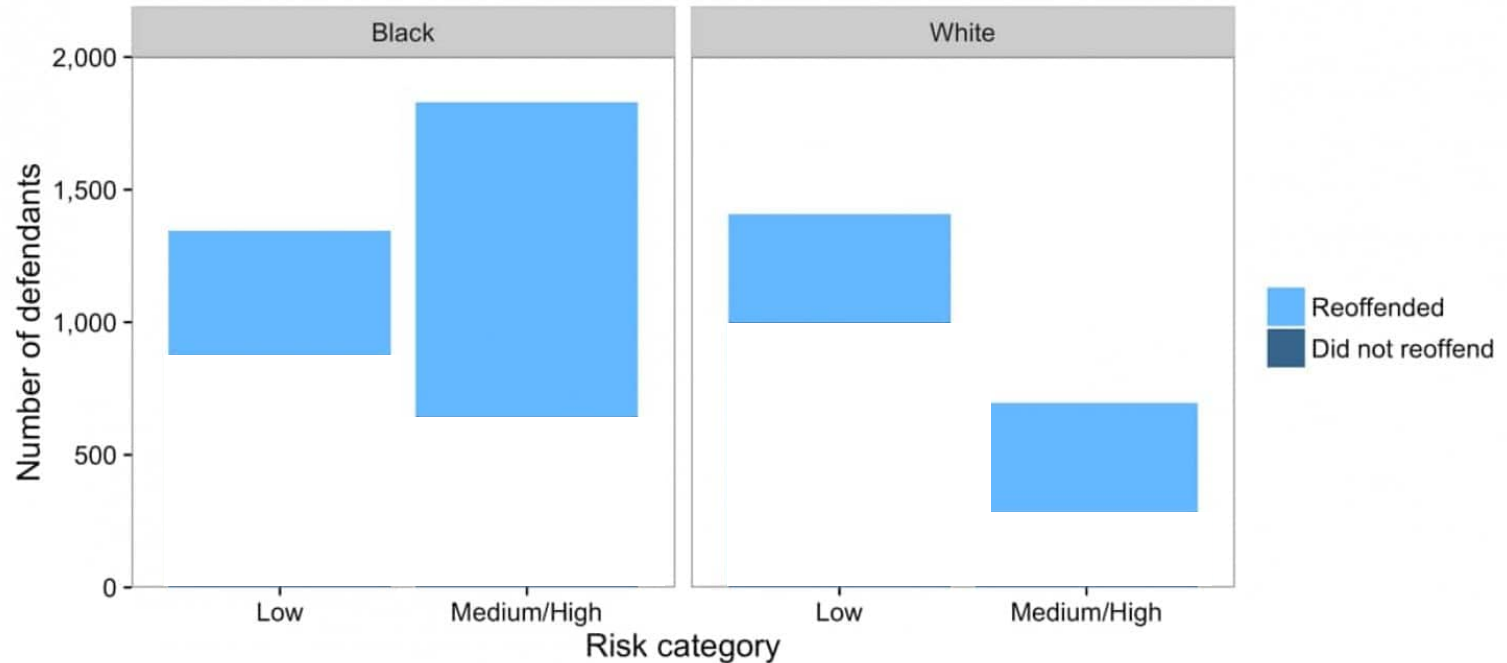


Propublica Argument (1)



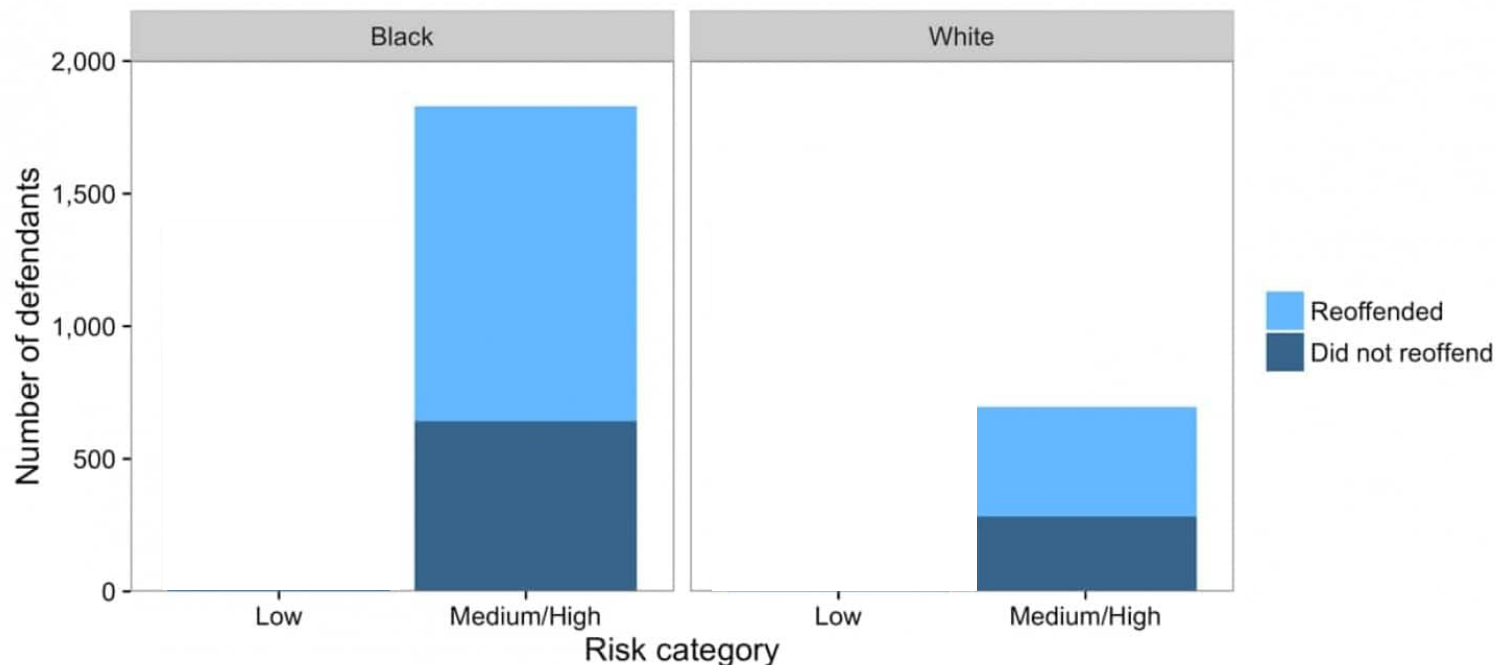
Black people who didn't reoffend far more likely than white people to have been rated as med/high risk

Propublica Argument (2)



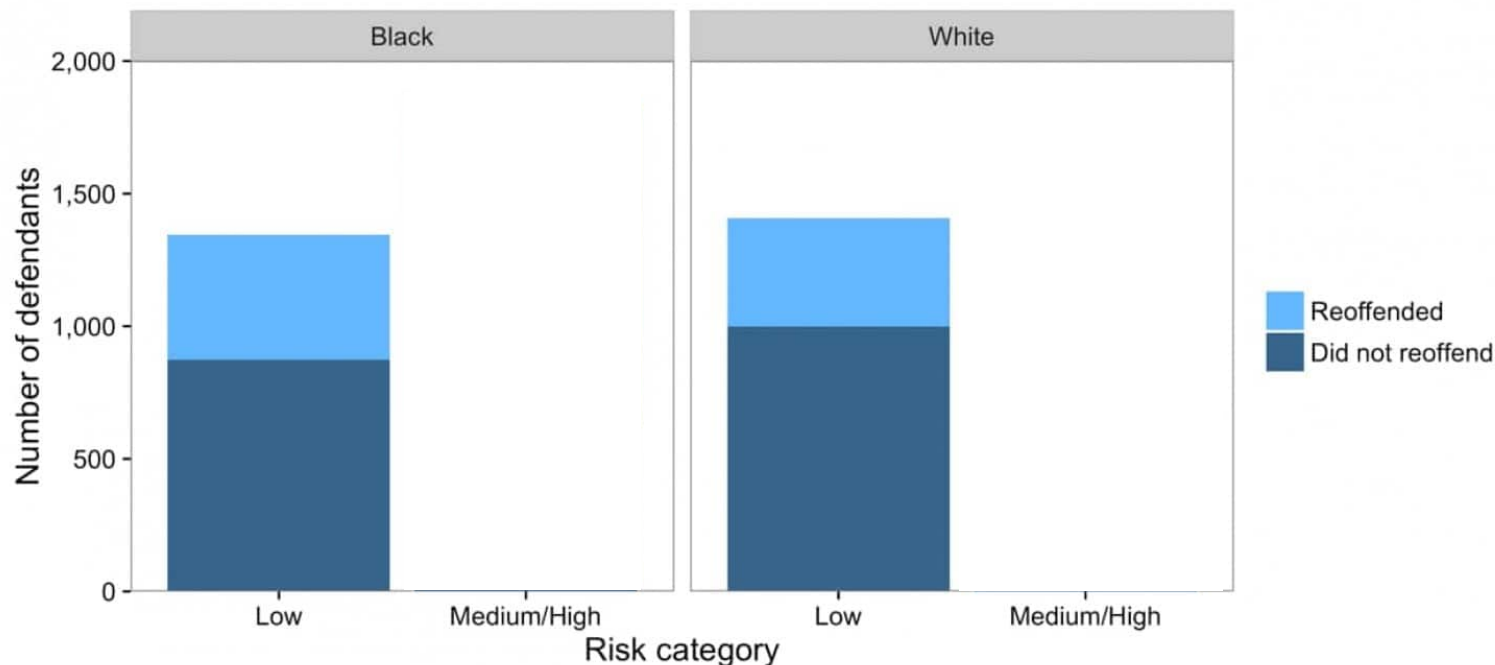
White people who reoffended far more likely than black people to have been rated as low risk

Northpointe Argument (1)



The chance of reoffending, for a med-high score, is the same for blacks and whites (scores mean the same thing)

Northpointe Argument (2)

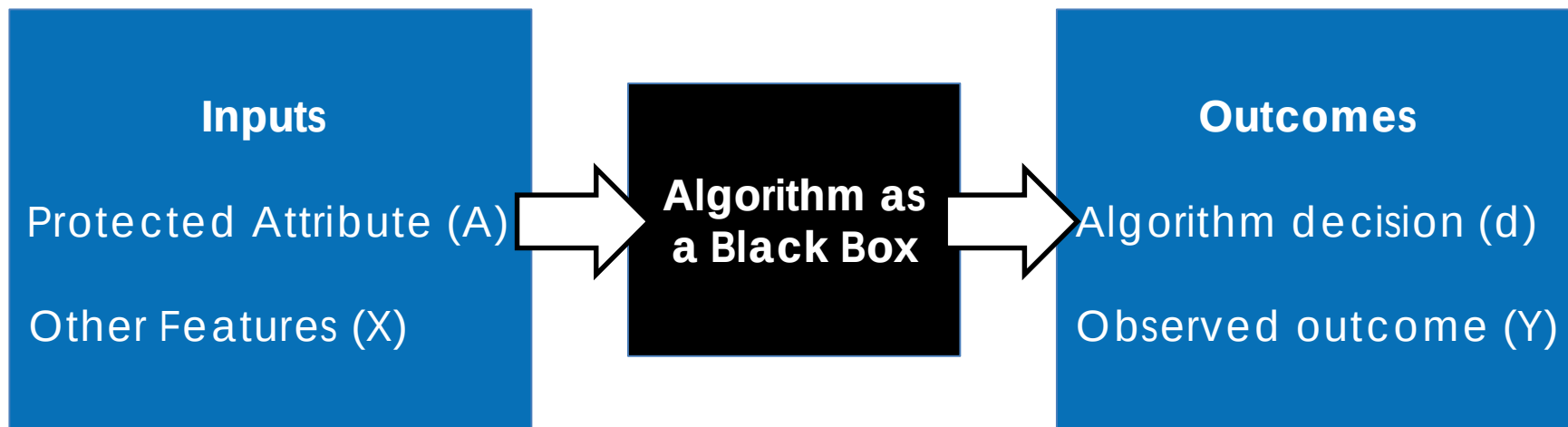


The chance of reoffending, for a low score, is the same for blacks and whites (scores mean the same thing)

Incompatible Definitions – What Should We Do?

- COMPAS inspired a lot of research into compatibility of fairness metrics
- It was found that for cases like this, where the outcome of interest (in this case recidivism) occurs at different rates across protected classes, you cannot simultaneously have more than two of:
 - False positive balance (first Propublica claim)
 - False negative balance (second Propublica claim)
 - Equal positive predictive values (Northpointe claims)(see Kleinberg et al., 2017, also Chouldechova, 2017)
- All are appealing ideals but we can't have all of them. Which one should be failed? How do we make such a tradeoff?

General Setting – Observational Metrics



- Many fairness concepts defined as property of joint distribution of $\{A, X, d, Y\}$
- Has advantage of needing no knowledge of the actual algorithm
- Some practical flaws but still widely used
- In the paper we also cover some causal concepts – we will ignore those today

Definitions of Fairness We Will Cover

Basis	Individual Fairness	Group Fairness
Equality of decision (d)	Unawareness	Demographic Parity
Equality of decision (d) conditional on another observed attribute (X or Y)		Conditional Demographic Parity Equalised Odds
Equality of observed outcome (Y) conditional on decision (d)		Calibration
Equality of decision conditional on unobserved “type”*	Actuarial Fairness	Actuarial Group Fairness

*Note - the concept of unobserved “type” will be introduced later

“Unawareness” – Just Don’t use the Protected Attribute!

- Reasonable initial thought on how to make things “fair”; widely applied in practice
- Ensure that for any individual, if we change the protected attribute and leave everything else the same we don’t change the decision:

$$d(x, A = a) = d(x, A = a') \text{ for all } x$$

- Often we do not have access to A: “unawareness” by default.
- Aligns to the common legal requirement to avoid direct discrimination; often mandated
- Widely acknowledged to pose risks of unintended indirect discrimination, due to the protected attribute commonly being “redundantly encoded” in other variables (e.g. as explored in Pedreschi et al. 2008)

“Demographic Parity” – A Quota System

- Demographic parity is perhaps a more stringent fairness criterion requiring that the “expected” or average decision is independent of A:

$$E(d \mid A = a) = E(d \mid A = a')$$

- For example if you are a political party or a company board, you might apply a quota system to your recruitment to ensure balance by gender
- In insurance, this is analogous to “community rating” – ensuring balance in the average premiums for particular groups irrespective of other characteristics (e.g. certain risk attributes).

Demographic Parity – Issues and Considerations

- What population should the averages be computed over? Is it:
 - the population of active customers,
 - members of the public who choose to quote / apply,
 - all residents over 18,
 - something else?
- Quota systems can cause harm. e.g. if we are issuing loans, what if one group is generally not creditworthy but the quota means we still have to issue loans?
 - The risks of various fairness criteria causing long term harm is being studied in more depth (e.g. see Liu et al. 2018 for an introduction).
- Note that demographic parity does not imply unawareness (and vice versa). It requires balance on average across groups, not for any one individual.
 - In particular, a quota system usually requires knowledge of the protected attribute to ensure the quota is met. Hence it may conflict with unawareness.

Conditional Demographic Parity

- Now we condition on some “legitimate” subset of factors Z:

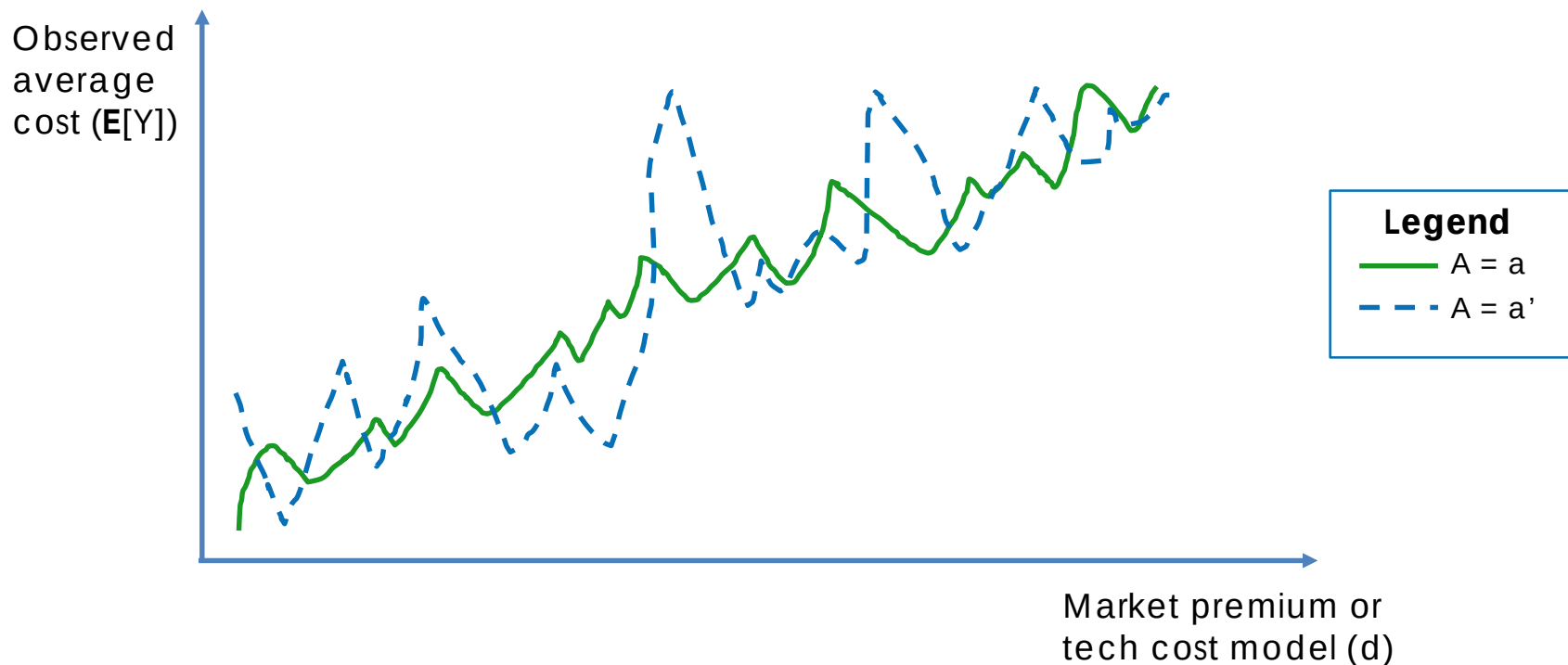
$$\mathbf{E}(d \mid Z, A = a) = \mathbf{E}(d \mid Z, A = a')$$

- So we are still ensuring equality of decisions, but over smaller “comparable” subsets
- Our choice of Z could be based on heuristic arguments around causal connection with Y and the degree of agency of an individual.
- In insurance pricing for instance, it might make some sense to condition on sum insured bands or excess levels, even if we desire “community rating”
- In general, we approach “unawareness” as we increase the granularity of Z.

“Calibration” – The Scores Mean the Same Thing

- Here, we ensure that for every decision we could take, the expected value of the observed outcome Y is equal across protected classes
 - i.e. $E(Y \mid d, A = a) = E(Y \mid d, A = a')$
- This was effectively Northpointe’s defence in the COMPAS example:
 - For those who were rated high risk (i.e. where $d = 1$), $P(Y = 1)$ was similar for blacks and whites
 - Similarly for those rated as low risk (i.e. where $d=0$)
- Many common model diagnostic plots are analogous to checking calibration

Calibration – Adapting a Common Diagnostic Plot



“Equalised Odds” – Balance the Errors

- Here we must make sure the expected decision $E(d)$ is independent of A , conditional on the outcome Y :

$$E(d \mid A = a, Y = y) = E(d \mid A = a', Y = y)$$

- In practice in binary classification this means the different forms of error (false positive, false negative) are balanced
- This was effectively Propublica’s argument in the COMPAS example
- So we already know that calibration and equalised odds are incompatible in most situations

Equalised Odds – Issues in Insurance Setting

Exposure 1:
10% chance
\$1,000 loss

Risk Type	Population Size		Risk Rated Prices (\$)
	A=a	A=a'	
Risk 1 – only Exp.1	100	150	100
Risk 2 – Exp 1 and 2	100	50	200

Exposure 2:
1% chance
\$10,000 loss

- Assume that Exposure 2 event didn't occur in your data
- Consider “non-claimers”. Higher average premium for A = a than for A = a', since A = a has more Exposure 2. Fails equalised odds
- A way to equalise this is to share cost of additional risk given by exposure 2 – we need demographic parity!

Actuarial Fairness

- We can avoid the issues in equalised odds by relating the decision to an unobserved attribute: the “true type” μ .
- “True type” represents a characteristic with most direct moral relevance to the decision process at hand. With it, the decision process should be deterministic.
 - In insurance pricing, μ can represent the true expected risk cost.
- If we set $d = \mu$, this is consistent with the traditional notion of “Actuarial Fairness”, perhaps first introduced by Kenneth Arrow

“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.”

“Uncertainty and the Welfare Economics of Medical Care”(Arrow,1963)

Actuarial Group Fairness

- The observed outcome Y may be considered to be a noisy (and potentially biased) realisation of the unobserved “true type” μ
- Thus we can adapt the notion of equalised odds into the notion of “**actuarial group fairness**”, defined as:

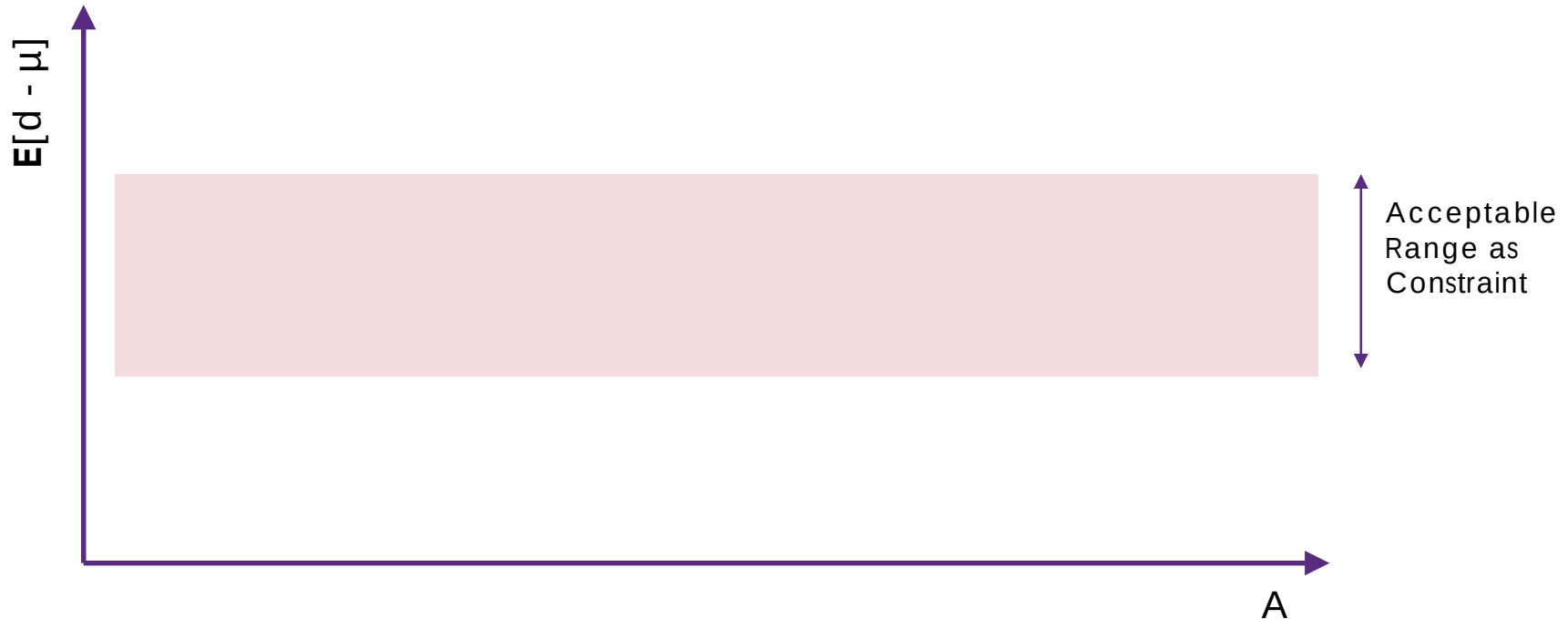
$$\mathbf{E}(d \mid \mu, A = a) = \mathbf{E}(d \mid \mu, A = a')$$

- This means we are equalising expected decisions, evaluated for each “true type”
- If Y corresponds to μ exactly, we recover the original definition of equalised odds. This would be true in situations where
 - an individual has perfect agency over the outcome Y , and
 - Y is perfectly observed.

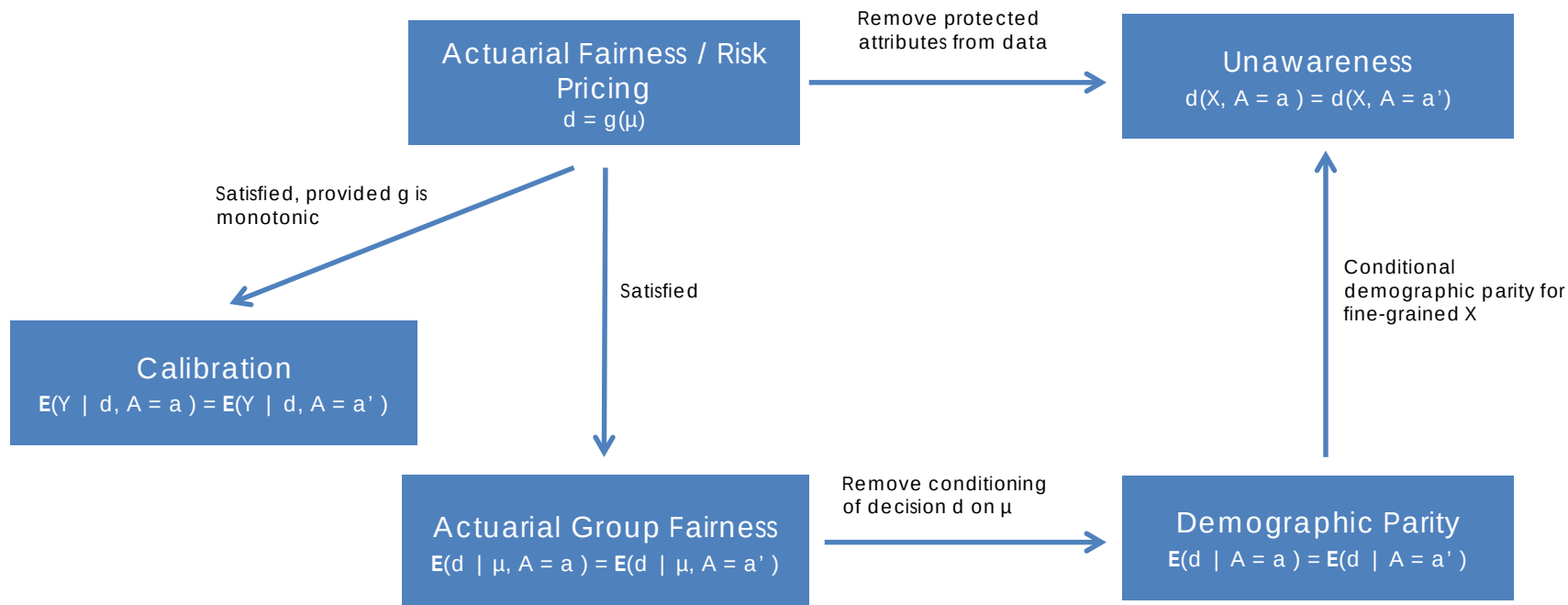
Actuarial Fairness - Opportunities

- Actuarial fairness has some considerable history, but requires judgement to apply in practice, since μ is unobserved
- Many techniques developed by actuaries to estimate μ from available insurance claims data, allowing decisions to directly reference an estimate for μ , rather than being solely reliant on raw data which may have known biases
- Similar logic could be applied in many other settings, adding judgement and professional ethics to algorithm design and data driven processes across many settings
- Actuarial fairness can be of direct relevance in many settings, particularly those where equalised odds may have conceptual challenges

Actuarial Group Fairness – Use as a Constraint



Recap and Relationships Between Concepts



A Reminder of Incompatibility

- As the COMPAS case illustrated, many intuitively desirable fairness concepts are incompatible in real world settings
- As we increase the number of fairness concepts, this problem grows rapidly. This presentation has only covered the more commonly seen concepts in the recent literature – there are dozens of others
- Important implications for any desire to create a generalised state of “fairness” – such decisions will require tradeoffs to be defined, made and justified
- Tradeoffs will likely be situational – not amenable to simple rules

So What Can You Do?

1. Create Internal Clarity

- Discuss what fairness means with relevant stakeholders
- Write down all the considerations
- Come to a decision
- Make everyone aware what that decision is

2. Acknowledge Imperfections

- Tradeoffs are unavoidable
- High level commitments to “act fairly” are not sufficient – we need to be more precise about what we actually mean, and acknowledge the trade-offs this entails.

3. Be Adaptable

- This is an evolving space
- Someone smart may well propose a better idea or metric tomorrow – if so, use it!
- Avoid staking too strong a claim on any “answer”, however appealing

4. Be Humble

- People genuinely disagree on the “right answer”
- Very rarely is someone “wrong”
- However strong our opinion, we may be called to change it

Get Involved!

- Society is having a complex, technically challenging debate about fairness, **right now**
- Recent examples, in Australia:
 - Australian Human Rights Commission project on human rights and technology (<https://tech.humanrights.gov.au>)
 - Federal government dept. of Industry, Innovation and Science – AI Ethics Framework (with Data61) (<https://data61.csiro.au/en/Our-Work/AI-Framework>)
- Actuaries have thought about fairness a lot over the years, in our particular domains, and can draw on some experience that may have broader applicability
- Actuaries have a good combination of technical skills and real world experience. We can and should contribute more to this debate