Criminally Incompetent Academic Misinterpretation of Criminal Data - and how the Media Pushed the Fake News

Norman Fenton[1] and Martin Neil[2]

19 January 2017

On 17 Jan 2018 multiple news sources (e.g. [1][2][3][4]) ran a story about a new research paper [4] that claims to expose both the inaccuracies and racial bias in one of the most common algorithms used by parole boards to predict recidivism (i.e. whether or not a defendant will reoffend).

The research paper was written by the world famous computer scientist Hany Farid (along with a student Julia Dressel).

But the real story here is that the paper's accusation of racial bias (specifically that the algorithm is biased against black people) is based on a fundamental misunderstanding of causation and statistics. The algorithm is no more 'biased' against black people than it is biased against white single parents [6], old people [7], people living in Beattyville Kentucky [8], or women called 'Amber' [9]. In fact, as we show below, if you choose ***any factor*** that correlates with ***poverty*** you will ***inevitably*** replicate the statistical 'bias' claimed in the paper. And if you accept the validity of the claims in the paper then you must also accept, for example, that a charity which uses poverty as a factor to identify and help homeless people is being racist because it is biased against white people (and also, interestingly, Indian Americans [10]).

The fact that the article was published and that none of the media running the story realise that they are pushing fake news is what is most important here. Depressingly, many similar research studies involving the same kind of misinterpretation of statistics repeatedly result in popular media articles that push a false narrative of one kind or another.

The algorithm (COMPAS) [11] being discussed by Farid and Dressel is exactly the kind that we have analysed with our colleagues in forensic psychiatry (see [12][13]). We know that such algorithms – which use only data to learn a regression model – have limited predictive accuracy because they are based on the data readily available rather than the data actually needed (in our studies we showed that the models can only be improved by incorporating expert judgment with causal relations and intervention factors that may be hard to measure). Farid and Dressel demonstrate this limited predictive accuracy. If the paper had restricted its claims to this limitation it would have been fine but not novel (especially as it failed to explore the issues of expert judgment, causality and interventions). The problem is that, in addition to this limitation, Farid and Dressel - along with journalists picking up on it –

[1] Professor of Risk, Queen Mary University of London (n.fenton@qmul.ac.uk)

[2] Professor in Computer Science and Statistics, Queen Mary University of London (m.neil@qmul.ac.uk)

We acknowledge support from ERC (Project ERC-2013-AdG339182-BAYES_KNOWLEDGE) and Leverhulme (Project CAUSAL-DYNAMICS)

emphasised that the COMPAS algorithm is also 'racially biased'. Indeed, this specific claim comes first in the introduction (our emphasis):

> *"..**the predictions were unreliable and racially biased**. COMPAS's overall accuracy for white defendants is 67.0%, only slightly higher than its accuracy of 63.8% for black defendants. **The mistakes made by COMPAS, however, affected black and white defendants differently**: Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at 28.0%. **In other words, COMPAS scores appeared to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants**."*

They say this despite acknowledging that the algorithm does ***not*** include race as one of its factors. They say that 'other aspects of the data may be correlated with race'. But it only needs one of the algorithm's factors to be 'correlated with race' for the above results to be inevitable. And we can prove this[3]. Let's think of the simplest case where an algorithm (let's call it SIMPLE) is based on just *one* factor, namely whether or not the defendant is ***poor*** (this is one of the COMPAS factors). And let's suppose SIMPLE works as follows:

> If a person is poor then predict "reoffend", otherwise predict "not reoffend"

Note that SIMPLE (like COMPAS) ***does not rely on – or use - any information about race***. From data, we know that a higher proportion of poor people reoffend than people who are not poor [14]. Then SIMPLE is, like the COMPAS algorithm, 'accurate' in the sense that, if all you know about a defendant is whether or not they are poor, it will be a better predictor of recidivism than just tossing a coin to decide.

Now it is known that black people are more likely to be poor than whites [10].

Hence, SIMPLE is more likely to predict that black people will reoffend than whites. It's as simple as that.

But, specifically, it also means that black people who do not reoffend are more likely to be predicted to reoffend than white people who do not reoffend. And also that white people who do reoffend are more likely to be predicted not to reoffend than black people who do reoffend. So, using the exact words that Farid and Dressel said about COMPAS**,** our algorithm SIMPLE:

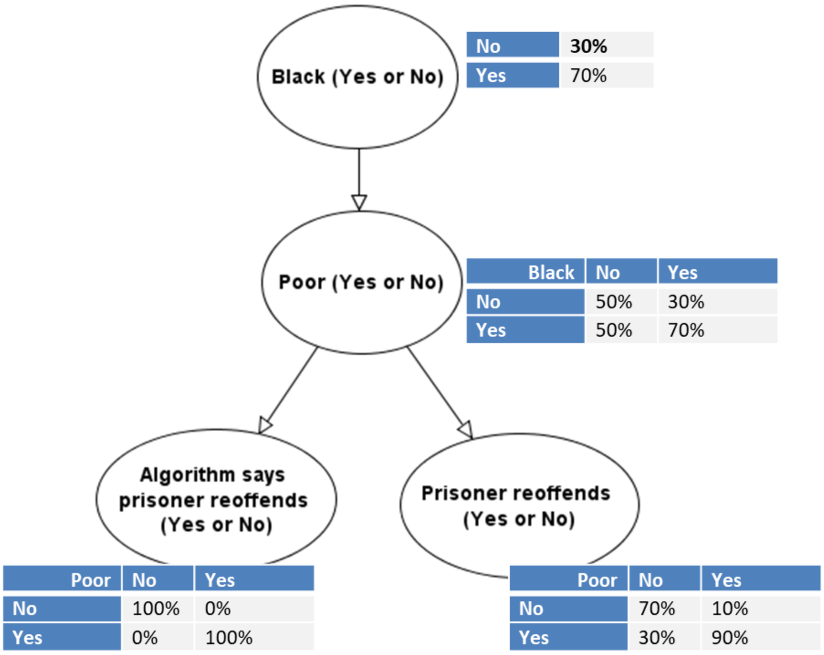
> favors white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants.

Clearly, the SIMPLE algorithm ***is biased against poor people.*** That does not make it either racist or useless**.** It is no more biased against black people than it is against

[3] Note that we do not claim to have analysed the full data, as we do not have access to it

any of the people with any of the other attributes we listed earlier that correlate positively with poverty.

A formal demonstration of our statistical argument can be made using the simple model and assumptions shown in Figure 1 (this is what is called a Bayesian network [15] – we use this to model the relationships between the factors and automatically compute the necessary probabilities[4]). The model's statistical assumptions are shown by the probability tables (for example, the table for the node "Poor" says that we assume 50% of people who are not black are poor and 70% of people who are black are poor).


Black (Yes or No)
Poor (Yes or No)
Algorithm says prisoner reoffends (Yes or No)
Prisoner reoffends (Yes or No)

No	30%
Yes	70%

Black	No	Yes
No	50%	30%
Yes	50%	70%

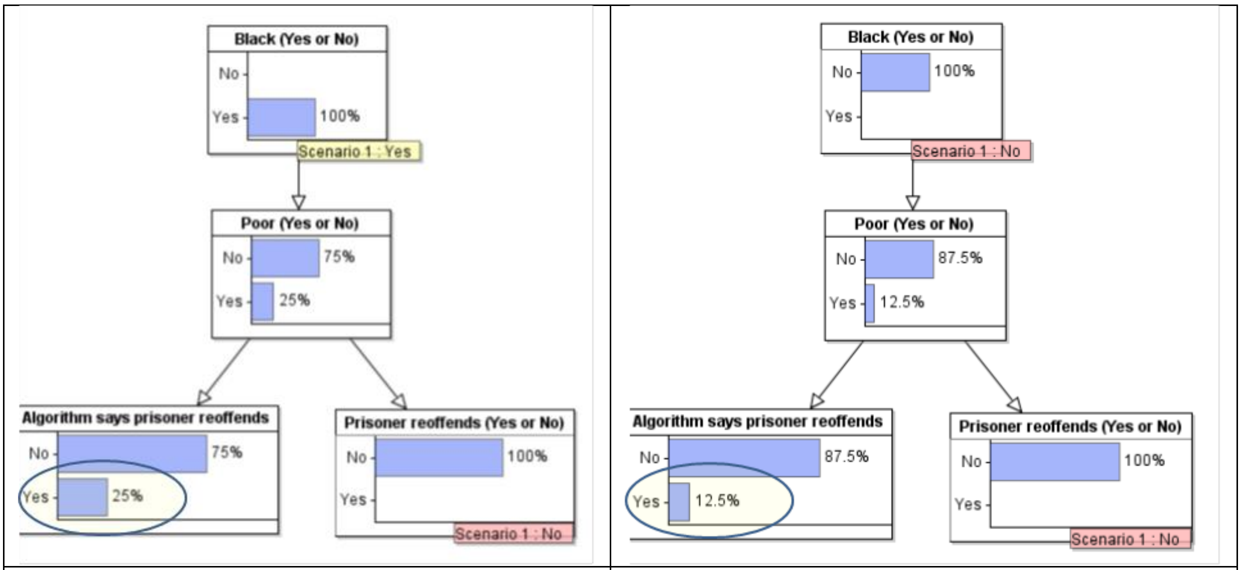
Poor	No	Yes
No	100%	0%
Yes	0%	100%

Poor	No	Yes
No	70%	10%
Yes	30%	90%

Figure 1: Model that captures the relationships and statistical assumptions

We can enter observations into any of the model nodes and it will calculate the correct statistical results for all of the unobserved nodes. This enables us to calculate the error rates for the algorithm as shown in Figures 2 and 3. Thus, in Figure 2a we see that Black defendants who do not reoffend are incorrectly predicted to reoffend at a rate of 25%, which is twice as high as their white counterparts (Figure 2b) at 12.5%; and in Figure 3a white defendants who do reoffend are incorrectly predicted to not reoffend at a rate of 25%, which is twice as high as their black counterparts (Figure 3b) at 12.5%. Note that this actually replicates the results of Farid and Dressel in the sense that the algorithm 'favours' white over blacks for both types of errors by 2 to 1.

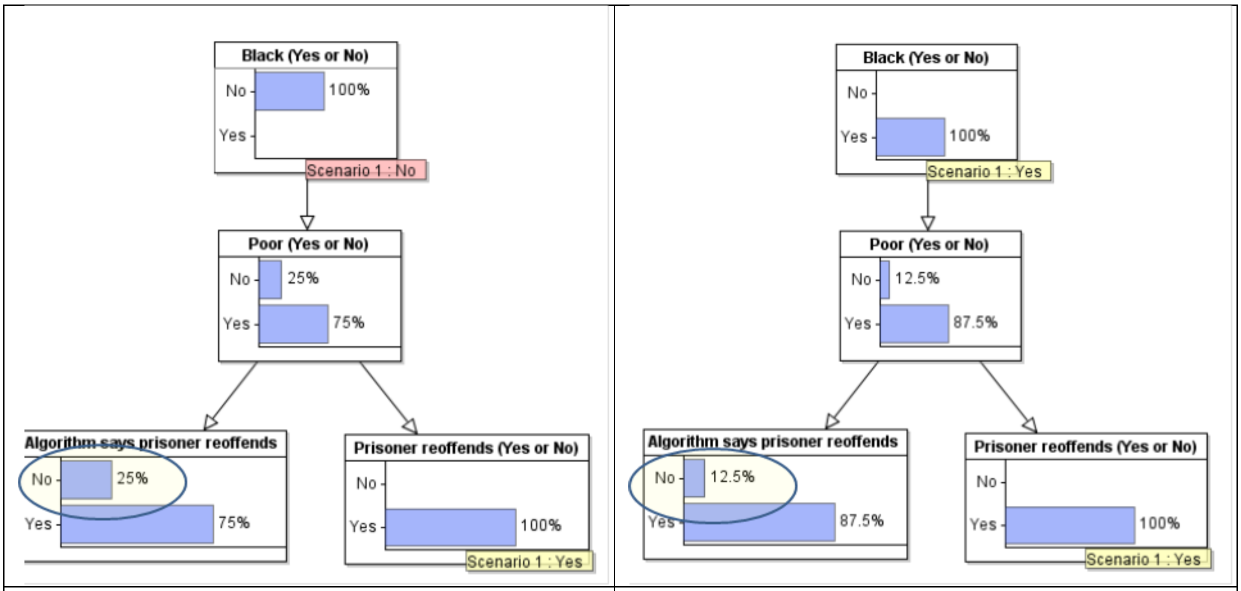
[4] These computations can be carried out by hand using Bayes Theorem in this case, but are rather laborious



(a) Black defendants: the error rate is 25%

(b) White defendants: the error rate is 12.5%

Figure 2: % of defendants who do not reoffend but who are incorrectly predicted to reoffend



(a) White defendants: the error rate is 25%

(b) Black defendants: the error rate is 12.5%

Figure 3: % of defendants who do reoffend but who are incorrectly predicted to not reoffend

To ram the point home about the futility and inevitably of the ‘racist’ claims we could – in our model - simply replace ***prisoners*** with ***people*** generally and “***reoffends”*** with “***shops in Walmart***”. By using only poverty as a factor, the algorithm is a reasonable predictor of whether or not a person shops at Walmart. The algorithm is no more or less racist than the COMPAS algorithm, and it is just as biased against black people as it is against white single parents. We could also replace “***reoffends***” with “***homeless***”; then, suppose the algorithm is used by a charity to predict people likely to become homeless (in order, say, to give them money). Because white people are less likely to be poor than black people, the algorithm is less likely to predict white people will become homeless than black people, and so white people would be less likely to get the money. So it would be just as ‘valid’ to state that the algorithm is biased against white people in this case as it is to claim the COMPAS algorithm is biased against black people.

The most desirable approach to developing algorithms such as COMPAS is to avoid socio-economic factors that correlate with race entirely, and move towards modelling morality and personal agency. Also, in [12][13] we showed how improved predictive accuracy is achieved by modelling causal relations and interventions, but this requires some expert judgment as the relevant data are not easily available. For example, the data did not contain information about interventions such as when the most 'at-risk' prisoners who were released were also closely monitored afterwards. Absence of information about this intervention explained a lot of the most serious inaccuracies in the model predictions. When it was added to the causal model the predictive accuracy improved, and the model became much more powerful and useful as a decision support tool. It also enabled what-if type analysis to be incorporated into decisions about whether to release a prisoner or not and so explore the sensitivity of decisions on particular circumstances.

It is hoped that parole boards using COMPAS treat people as individuals and not labels, and that they already take account in their decision-making of the factors (many of which are already in COMPAS) that may be related to personal agency rather than just socio-economic factors. It is actually possible that these other factors might measure such things as moral disposition, yet give the same statistical results. Given this it may well be that it is the model detractors who are fixated on race and not the parole boards.

[1] https://www.theguardian.com/us-news/2018/jan/17/software-no-more-accurate-than-untrained-humans-at-judging-reoffending-risk

[2] http://flip.it/nE3dB7

[3] https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/

[4] https://phys.org/news/2018-01-court-software-accurate-web-survey.html

[5] Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. Sci. Adv. 4, eaao5580 (2018). http://advances.sciencemag.org/content/4/1/eaao5580.full

[6] https://www.gingerbread.org.uk/policy-campaigns/publications-index/statistics/

[7] https://www.ncoa.org/news/resources-for-reporters/get-the-facts/economic-security-facts/

[8] http://money.cnn.com/2017/02/06/news/economy/donald-trump-beattyville-kentucky/index.html

[9] https://www.theatlantic.com/magazine/archive/2005/06/baby-names/303974/

[10] https://en.wikipedia.org/wiki/List_of_ethnic_groups_in_the_United_States_by_household_income

[11] COMPAS manual: http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf

[12] Constantinou, A., Freestone M., Marsh, W., Fenton, N. E. , Coid, J. (2015) "Risk assessment and risk management of violent reoffending among prisoners", Expert Systems With Applications 42 (21), 7511-7529. http://dx.doi.org/10.1016/j.eswa.2015.05.025

[13] Constantinou, A. C., Yet, B., Fenton, N., Neil, M., & Marsh, W. (2016). Value of Information analysis for interventional and counterfactual Bayesian networks in forensic medical sciences. Artificial Intelligence in Medicine. 66, pp 41-52 doi:10.1016/j.artmed.2015.09.002

[14] Kubrin, C. E. & Stewart, E. A. Predicting Who Reoffends: The Neglected Role of Neighbourhood Context in Recidivism Studies Criminology 44, 165–197 (2006).

[15] Fenton, N. E. & Neil, M. Risk Assessment and Decision Analysis with Bayesian Networks. (CRC Press, 2012).