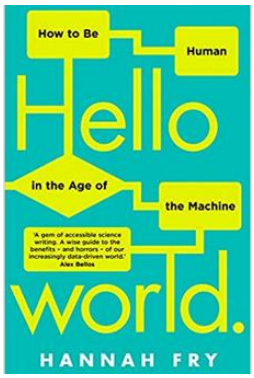
Hannah Fry's "Hello World" and the Example of Algorithm Bias

Norman Fenton, 16 March 2019



1 Introduction

"Hello World" (Fry, 2018) is an excellent book by Hannah Fry that provides lay explanations about both the potential and threats of AI and machine learning algorithms in the modern world. It is filled with many excellent examples, and one that is especially important is in Chapter 3 ("Justice") about the use of algorithms in the criminal justice system. After highlighting concerns about an algorithm called COMPAS (Northpointe, 2015) that is used to determine the risk of people reoffending, Hannah provides a simple theoretical example of an algorithm to determine the risk of a person being a murderer. The example demonstrates the extremely important point that there is an inevitable trade-off between 'accuracy' and 'fairness' when it comes to algorithms that make decisions about people; essentially that you can have algorithms that are accurate or fair but not both. In the example the algorithm is 'unfair' to men because they are more likely to be wrongly identified as high risk.

However, while the overall thrust and conclusions of the example are correct the need to keep any detailed maths out of the book might leave careful readers confused about whether the example really demonstrates the stated conclusions. I feel it is important to get the details right because the issue of algorithmic fairness is of increasing importance for the future of AI, yet is widely misunderstood. For example (Fenton et al., 2018a) highlights fallacies relating to widespread criticism of the COMPAS algorithm (Northpointe, 2015) that Hannah uses to motivate her example.

In this short report I will present the example as described by Hannah in Section 2. In Section 3 I will explain what is missing from the example, namely any explicit calculation of the ***false positive rates*** of the algorithm and I will also explain the danger of making a mistake called the ***fallacy of the transposed conditional***. In Section 4 I will show how Bayes theorem (and some other assumptions) are needed to compute the false positive rates for men and women. Finally, in Section 5 I will show why and how a causal model of the problem (namely a Bayesian network model) makes everything much clearer.

2 The example

The example is stated as follows, with the two pictures referred to shown here in Figure 1.

> *"Imagine stopping people in the streets and using an algorithm to predict whether each person will go on to commit homicide. Now, since the vast majority of murders*

are committed by men (in fact, worldwide 96% of murderers are male), if the murderer-finding algorithm is to make accurate predictions it will necessarily identify more men than women as high risk. Let's assume our murder detection algorithm has a prediction rate of 75% that is to say three quarters of the people the algorithm labels as high risk are indeed Darth Vaders.

Eventually after stopping enough strangers you have 100 people flagged by the algorithm as potential murderers. To match the perpetrator statistics, 96 of those 100 will necessarily be male, 4 will be female. There is a picture below to illustrate. The men are represented by dark circles, the women shown as light grey circles.

Now since the algorithm predicts correctly for both men and women at the same rate of 75%, one quarters of the females, and one quarter of the males will really be Luke Skywalkers: people who are incorrectly identified as high risk, when they don't actually pose a danger.

Once you run the numbers, as you can see from the second image here, more innocent men than innocent women will be incorrectly accused just by virtue of the fact that men commit more murder than women.

Hannah concludes:

"The outcome is biased because reality is biased. More men commit homicides, so more men will be falsely accused of having the potential to murder. Unless the fraction of people who commit crimes is the same in every group of defendants, it is mathematically impossible to create a test which is equally accurate at prediction across the board and makes false positive and false negative mistakes at the same rate for every group of defendants".

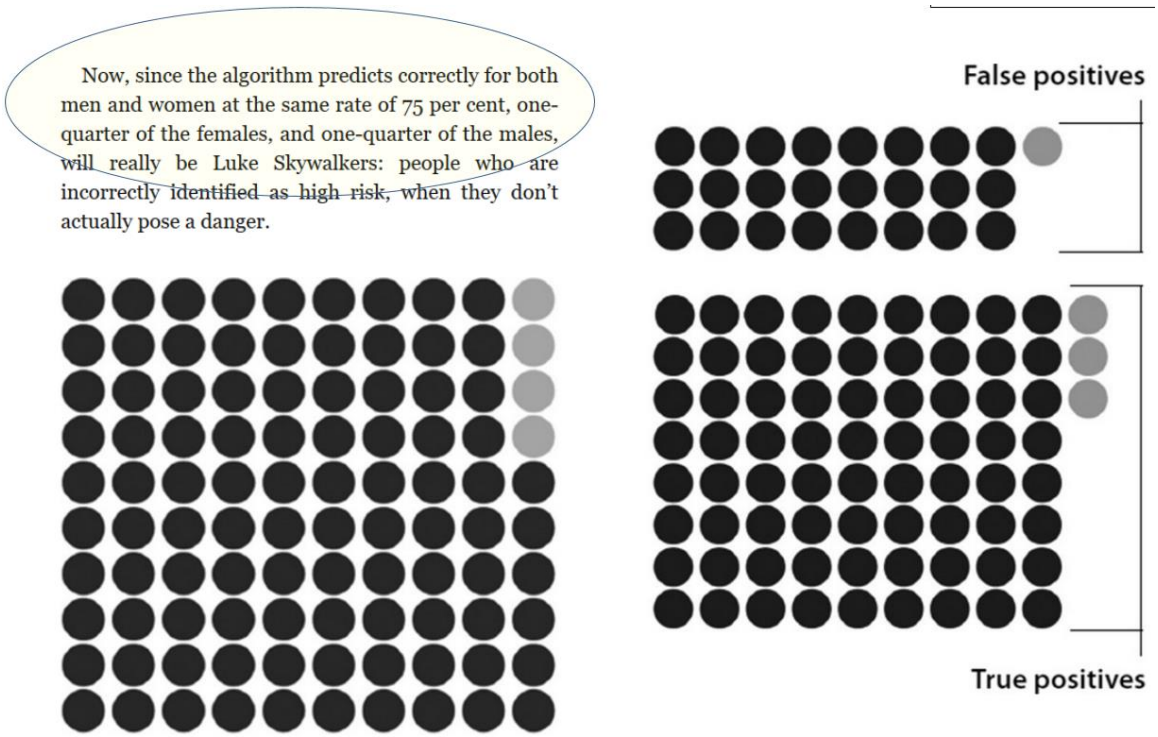


Figure 1 Hannah's example with a key assumption highlighted

3 What is missing from the example

The key thing missing from the example and its explanation is that it provides no indication about the thing we are ultimately most interested in, namely **the false positive rates**. Specifically:

> ***The false positive rate is the probability the algorithm will wrongly identify a non-murderer as high risk.***

In particular, we want to know the separate false positive rates for men and women, as the key implication is that the former is higher than the latter.

To determine these rates, we need to make a number of assumptions that are either missing from the example or are only implicit, and only then can we apply Bayes theorem to get the result we need. First let's introduce some notation.

> Let *M* be the assertion "*Person is a murderer*", and $\overline{M}$ its negation (i.e. "*Person is a non-murderer*")
>
> Let *X* be the assertion "*Algorithm reports positive (high risk)*", and $\overline{X}$ its negation (i.e. "*Algorithm reports negative*")

What we are told is that the probability of *M* given *X* is equal to 75%. We write this as the probability statement.

$$P(M|X) = 0.75$$

It follows that

$$P(\overline{M}|X) = 0.25$$

(since these two probabilities must sum to 1). This is the reasoning used in the example to explain why, out of 100 people who the algorithm rates as high risk, 75 will be murderers and 25 non-murderers.

But what we are most interested in are the following probabilities:

$P(X|M)$ (this is the ***true positive rate*** for the algorithm; the ***false negative rate*** is one minus this, i.e. $P(\overline{X}|M) = 1 - P(X|M)$)

$P(X|\overline{M})$ (this is the ***false positive rate*** for the algorithm; the ***true negative rate*** is one minus this, i.e. $P(\overline{X}|\overline{M}) = 1 - P(\overline{X}|\overline{M})$)

Moreover, we want to break these down into error rates for males and females.

One of the most common mistakes people use when reasoning about probability is the so-called ***transposed conditional fallacy***. This is to assume that the probability of a statement like A given B is always the same as the probability of B given A and that we could therefore deduce from above that:

$$P(X|M) = P(M|X) = 0.75$$

$$P\left(X\middle|\overline{M}\right) = P\left(\overline{M}\middle|X\right) = 0.25$$

In fact, it turns out that, while the first equality (the true positive rate) ***does*** hold in this example (once we introduce some reasonable assumptions) the second (the false positive rate) certainly does not.

In general, the correct relationship between two transposed conditional probabilities is provided by Bayes' theorem, which for the false positive, asserts:

$$P(X|M) = \frac{P(M|X) \times P(X)}{P(M)}$$

The reason, why in this case, $P(X \mid M)=P(M \mid X)$ is because it is reasonable to assume that $P(X)=P(M)$. This is because in the example Hannah asserts that the algorithm "*preserves the perpetrator statistics*". This means that, in addition to the fact that it preserves the 96% male proportion in its positive outcomes, we can also assume that, whatever the overall population proportion of murderers, $P(M)$, the proportion of positive outcomes produced by the algorithm, $P(X)$, will be the same. For example, if $P(M)$ is, say, 1 in 1,000 then $P(X)$ will also be equal to 1 in 1000. So, in this special case the 'transposed conditionals' are equal and we can conclude that the true positive rate is 75%. Moreover, since we also know that $P(X|M)$ is the same for men and women, we can also conclude that the true positive rates for both men and women are 75%. This also means the false negative rates for men and women are both equal to 25%.

However, ***things are very different for the false positive rate*** which, by Bayes, is computed as:

$$P\left(X\middle|\overline{M}\right) = \frac{P\left(\overline{M}\middle|X\right) \times P(X)}{P(\overline{M})}$$

Let us assume, as above that $P(M) = P(X) =1$ in 1,000. Then:

$$P\left(X\middle|\overline{M}\right) = \frac{0.25 \times 0.001}{0.9999} \approx 0.00025$$

So, the overall false positive rate is very low, about 1 in 4,000. This means that for every 100,000 people about 25 would be falsely rated as high risk by the algorithm. Note that this key information was missing from Hannah's example. But that is the overall false positive rate for all people. We next turn to the separate false positive rates for men and women.

4 Calculating the false positive rates for men and women

To calculate the different false positive rates for men and women we need to know the respective non-murderer probabilities for men and women as well as the probabilities of these given X. Let us denote:

a. *men who are murderers*
b. *men who are non-murderers*
c. *women who are murderers*
d. *women who are non-murderers*

Assuming that there are an equal number of men and women in the population, then for every 100,000 people there will be 100 murderers of whom 96 are men and 4 are women. Hence:

- $P(a)$ = 96/100,000 = 0.00096
- $P(b)$ = 499,904/100,000 = 0.49904
- $P(c)$ = 4/100,000 = 0.00004
- $P(d)$ = 49996/10,000 = 0.49996

To compute $P(b \mid X)$ we note the following from the example:

$$P(a|X) + P(b|X) = 0.96$$

(since we are told that 96% of people who the algorithm says are high risk are men).

But we also know that

$$\frac{P(a|X)}{P(a|X) + P(b|X)} = 0.75$$

(since we are told that 75% of men who the algorithm says are high risk are murderers).

From these two equations we compute that:

$P(a \mid X) = 0.24$

$P(b \mid X) = 0.72$

So, noting that b represents the non-murderer males, we can now compute the false positive rate for men:

$$P(X|b) = \frac{P(b|X) \times P(X)}{P(b)} = \frac{0.24 \times 0.001}{0.49904} = 0.000480923$$

That is a rate of less than 1 in 2000. It means that for every 100,000 men about 48 would be falsely determined as high risk by the algorithm.

Similarly, for women we note that

$$P(c|X) + P(d|X) = 0.04$$

and

$$\frac{P(c|X)}{P(c|X) + P(d|X)} = 0.75$$

From these two equations we compute that:

$P(c \mid X) = 0.03$

$P(d \mid X) = 0.01$

So, noting that *d* represents the non-murderer females, we can now compute the false positive rate for women:

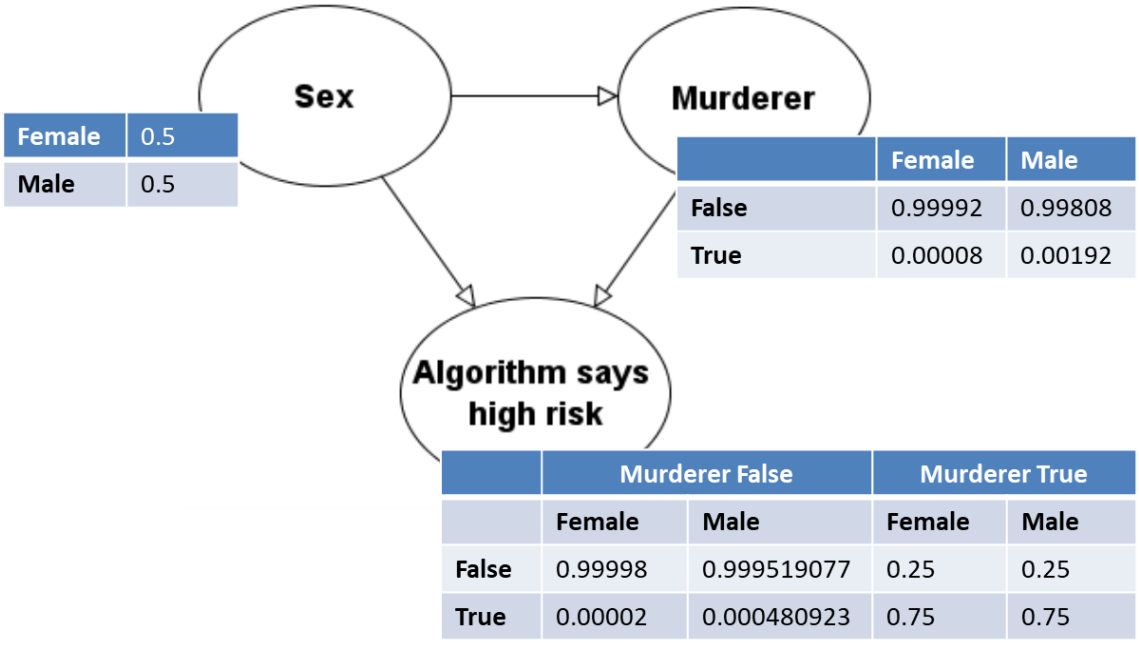
$$P(X|d) = \frac{P(d|X) \times P(X)}{P(d)} = \frac{0.01 \times 0.001}{0.49996} = 0.0000200016$$

That is a rate of less than 1 in 50,000. It means that for every 100,000 women about 2 would be falsely rated as high risk by the algorithm.

In summary we have computed the exact false positive rates for men and women and have confirmed that the algorithm is indeed 'biased' against men in the sense that the false positive rates for men are higher than for women. This is, indeed an inevitable outcome of the requirement for the algorithm to be 'accurate' in the sense that was described.

5 A causal model of the problem

As suggested convincingly by Pearl (Pearl et al., 2018) these types of problems become much clearer to analyse once we construct a causal model that makes explicit the underlying variables and their relationships. The appropriate causal model is shown in Figure 2.



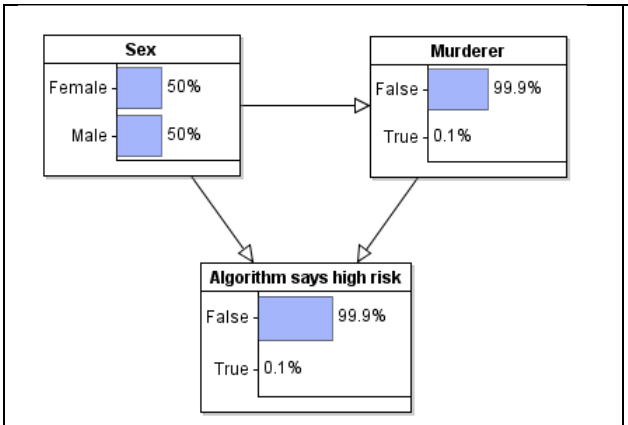
Female	0.5
Male	0.5

	Female	Male
False	0.99992	0.99808
True	0.00008	0.00192

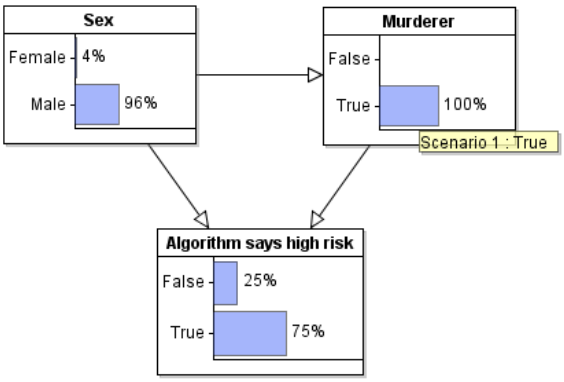
	Murderer False		Murderer True	
	Female	Male	Female	Male
False	0.99998	0.999519077	0.25	0.25
True	0.00002	0.000480923	0.75	0.75

Figure 2 Model with all probability parameters shown

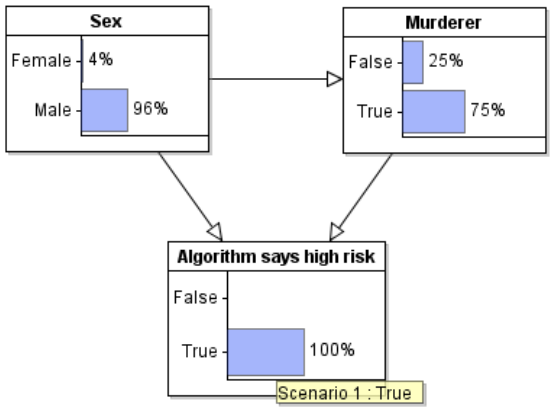
The model is an example of a Bayesian network because we have added the necessary tables showing the conditional probability values associated with each variable. The table values for the variable "Algorithm says high risk" simply use the results of the computations in Sections 3 and 4. The table values for the variable "Murderer" uses Bayes theorem (since we know P(Sex | Murderer), P(Murderer) and P(Sex) we can compute and P(Murderer | Sex)). With these table values we can run the model using a Bayesian network tool (Agena Ltd, 2019) and check the results as shown in Figure 3. Note that all of the required assumptions are preserved.



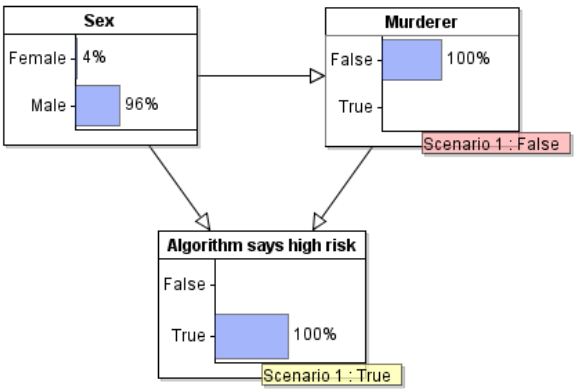
a) Model in its prior state (no observations). Note that both the murderer rate and the probability the algorithm says high risk are 1 in 1000



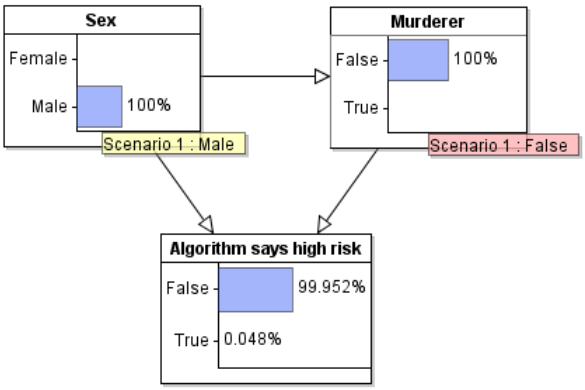
c) Confirming that 96% of murderers are males and that the true positive rate is 75%



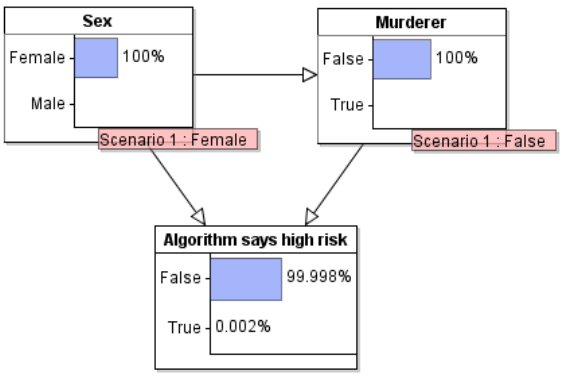
c) When we observe algorithm given a positive result we see that there is 96% chance it will be a male and there is a 75% chance the person really is a murderer.



d) When the algorithm is known to provide a false positive there is a 96% chance the person is a male.



e) Probability algorithm gives false positive for men



f) Probability algorithm gives false positive for women

Figure 3 Running the model with different observations

More details on how to build and run causal Bayesian network models can be found in (Fenton et al., 2018b).

References

Agena Ltd. (2019). AgenaRisk. http://www.agenarisk.com. Retrieved from http://www.agenarisk.com

Fenton, N. E., & Neil, M. (2018a). Criminally Incompetent Academic Misinterpretation of Criminal Data - and how the Media Pushed the Fake News. https://doi.org/10.13140/RG.2.2.32052.55680

Fenton, N. E., & Neil, M. (2018b). *Risk Assessment and Decision Analysis with Bayesian Networks* (2nd ed.). CRC Press, Boca Raton.

Fry, H. (2018). *Hello world : how to be human in the age of the machine*. New York: W. W. Norton & Company, Inc.

Northpointe. (2015). Practitioner's Guide to COMPAS Core. Retrieved from www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect*. New York: Basic Books.