BOOK REVIEW - WEAPONS OF MATH DESTRUCTION 2016 - CATHY O'NEIL, DATA SCIENTIST

REVIEWED BY ANNELIES MOENS

I came across Cathy O'Neil at the IAPP Academy's annual conference in San Jose in September 2016. Cathy was a keynote speaker and had just released her book, "Weapons of Math Destruction". Her presentation struck a nerve with the audience, because by the time I was at the queue to buy her book, it had already sold out. Nevertheless, I got hold of a couple of advance copies from Amazon prior to its Australian release late last year.

I have been recommending this book as a must read and it seems to have found its way into the popular press here in Australia, with two articles I came across recently quoting from it. One was in the Sydney Morning Herald, "How Centrelink unleashed a weapon of math destruction", 8 January and another in the Australian Weekend Financial Review, "Oh, algorithms, you don't know me at all", 28 December 2016 – 2 January 2017. Also, the Association for Computing Machinery US Public Policy Council (USACM) released a "*Statement on Algorithmic Transparency and Accountability*" on 12 January 2017 – focusing on steps to prevent algorithmic bias.

As such, I thought it was timely to give Cathy O'Neil's book a review for iappANZ members.

Contrary to popular notion, not all data scientists are big data evangelists. Cathy O'Neil is a data scientist, has a PhD in Maths from Harvard, taught at Barnard College, worked for a hedge fund DE Shaw and worked at various startups building models that predict people's purchases and clicks. Based on her experience, the author looks at the dark side of big data.

What is a Weapon of Math Destruction (WMD)?

Firstly, the title of the book lends itself to an unforgettable name and conjures up the nasty impacts of weapons of *mass* destruction in the physical world. Here is how the author classifies models and algorithms that produce automated decisions as WMDs:

WMDs	Opposite of WMDs
- Opaque (we are modelled as shoppers, patients, loan applicants etc) – we see little of that modelling - Unregulated - Uncontestable (without feedback a statistical engine can continue with faulty and damaging analysis while never learning from its mistakes) - Have massive scale and damage.	- Transparent - Controlled by user and personal - Shared objectives - Updated - Auditable - Accountable

WMDs encode human prejudice, misunderstanding and bias and often punish individuals who happen to be the exception. People think that because a machine made the decision it can't be biased – well that is a fallacy for a WMD.

Real life examples – WMDs

Being an American, most of the author's examples are US based – it would be great to see some global examples from different cultures. The author starts the book with the story of Sarah Wysocki which made the press in 2012 and about whom journalists wrote extensively. Sarah was a 5th grade school teacher, who was dismissed as a bad teacher based on a WMD. She had excellent reviews from principals, students and parents. A new scoring system called IMPACT was launched to "measure" teaching effectiveness. IMPACT measured the educational progress of students and then calculated how much of their advance or decline could be attributed to the teacher. Many complex variables effectively led to random outcomes. Sarah was fired and the next day hired by a private school.

Another example is University rankings generally that lead to perverse incentives and outcomes. In particular, the author focused on the US News Best Colleges Ranking example. Rankings force everyone to shoot for the same goals, which creates a rat race and lots of harmful unintended consequences. In that particular example, she writes about the cost of tuition not being included – hence a contributing factor to skyrocketing costs of education in the USA. Since then, according to the author the US Department of Education has intervened to try to address this problem.

The prison and police systems also offer interesting WMDs. The author highlights the models that determine the risk of recidivism (re-offending) in determining jail sentences for convicted persons. She argues that the model itself contributes to a toxic cycle and helps to sustain it. The algorithms use proxies – like what the family of convicted persons do, where they live, previous offences, people like them, etc. This is where "people like you" that are deemed to be "you" has serious consequences. Price optimization is another area in which WMDs are used. The author provides an example from AllState Insurance which offered discounts of 90% off the average rate to 80% increases. It massively varied the cost of premiums, based on analyzing consumer and demographic data to determine the likelihood that consumers would shop for lower prices. A consumer deemed less likely to shop for lower prices would be charged more. The author states that every person had a different experience, and the models were optimized to draw as much money as they could from the desperate and ignorant.

Social media platforms and search engines are another minefield for WMDs. Platforms, like Facebook make judgments about which friends see what posts. They determine what we see and learn, much like search engines. These platforms are massive, powerful and opaque. As an aside, a new blog on the harmful privacy policies of Facebook was released in the New Year and writes about getting off Facebook to save your relationships with your friends.

Studies into search engines have tested undecided voters to see if they can be swayed in a particular direction depending on the search results they see. The author refers to a study into undecided voters in the USA and India, where these individuals were asked to use search engines to learn about upcoming elections. The search engines they used were programmed to skew search results, favouring one party over another. Those results shifted voting preferences by 20%. Pew Research shows that 73% of Americans believe search results are "accurate and trustworthy", which in my view would place massive accountabilities on those providers (much like the press). Indeed, Angela Merkel complained about the lack of search engine transparency endangering debate in The Guardian late last year.

Good examples of automated algorithms/models

It would have been remiss of the author had she not included what a good automated algorithm/model looks like. Here, the author provides the well-known example of baseball (think about the movie Moneyball).

In baseball, predicting who will win, understanding how to put a winning team together or how to play your best against upcoming opposition can all be fed by data. These models are generally:

- *Transparent* - fans, players, managers have access to the statistics (home runs, strikeouts) and can more or less understand how they are interpreted
- *Continually updated* - new games are fed in and mistakes to models corrected
- *Relevant* - Assumptions and conclusions are transparent ie: data is highly relevant to the outcome that is trying to be predicted
- *Direct sources* - Statistics come from direct sources, not proxies: ie: actual footage of games and scores
- *Shared Objectives* - People being modelled understand the process and share the model's objectives, for example winning and predicting baseball games

Lessons learnt

The top takeaways, in my view, from the book are as follows:

- Human decision making, while often flawed, has one chief virtue – it can evolve (it will be interesting to see how effective machine learning algorithms become)
- We have to explicitly embed better values into our algorithms, creating big data models that follow our ethical lead
- Standards can protect companies that want to do the right thing – because their competitors have to follow the same rules
- Like doctors, data scientists should pledge a Hippocratic Oath, one that focuses on the possible misuses and misinterpretations of their models
- Today the success of a model is often measured in terms of profit, efficiency or default rates - fairness and common good resist quantification. As such it is necessary to impose human values on these systems, even at the cost of profit or efficiency.
- The impact of models needs to be measured and audits of algorithms need to be conducted and feedback loops built in. Models must deliver transparency, disclosing the input data they're using as well as the results of their targeting.