The Perils of Prediction: Lessons for Regulators in the Age of Big Data

December 10, 2018

You've likely read one too many articles that start by telling you that big data is watching you and predicting what you will do or buy next. About how it knows if you are pregnant, or that sales of strawberry pop tarts surge before a hurricane. We'll leave it to others to explain the how and why of that. Our focus will be on the growing emphasis on the use of data and numbers to make predictions for the purpose of public protection. We want to flag for regulators some of the key legal and ethical issues relating to statistical prediction.

The first step is to identify the key issues. We will have to do this indirectly. Most of our regulatory clients oversee occupations or professionals of one sort or another. Statistical prediction is not as far along here as in other sectors. So we will take a tour through some other sectors that have either embraced predictive methods or where compelling issues have emerged with respect to them. The first place we will look is policing.

Policing

"In our society, we have no major crimes but we do have a detention camp full of would-be criminals." These words come from a character in Philip K. Dick's short story "The Minority Report". More people are likely familiar with the 2002 film of the same name starring Tom Cruise. The story and movie depicts a system that uses mutant psychics to predict crime. "Pre-crime" police units then use the predictions to arrest people to prevent a crime before it happens.

The media uses the label "minority report policing" to refer to the growing use by police of data and predictive methods. The more conventional name for this is predictive policing. It refers to the use of statistics and software to forecast crime and predict who will commit it. No psychics are involved. Instead, computer algorithms churn through data to forecast who is at risk to break the law or which areas will have a high crime rate. To some extent it is like minority report because police use predictions to intervene to prevent crime.

The crime forecasts may relate to different things. One approach is to focus on the person who is at risk to commit a crime. The most famous, or infamous, example is the Chicago police departments Strategic Suspect List, otherwise known as the "heat list". The idea behind the heat list is that a small percentage of young people account for most gun violence. If the police know the people who make up that percentage the police could intervene in the lives of those most at risk to commit or be a victim of gun violence.

The heat list uses an algorithm to assign a risk score to a person based on criminal record. The higher the score the greater the risk you will be involved in gun violence. Those with high scores receive a "custom notification." This involves a home visit from the police and a social worker and an influential member of the community, for example a pastor or sports coach. There is the offer of a referral to social services should the person wish to seek help in turning his or her life around. The visit includes the delivery of a letter setting out what the police know about a person's criminal past and why he or she is at risk. The letter provides a warning to the person that he or she will face the highest possible charges if he or she is arrested in the future.

The heat list stumbled in its early versions. A 2016 study by the Rand Corporation found that the list had problems with accuracy and had little effect on violent crime. Police were using it more as a "most wanted" list. Those on the list had no greater risk than a comparison group for involvement in gun violence. But being on the list brought a higher risk of arrest. A more recent version of the list may have higher accuracy. The Rand study related to the 2013 list. Chicago police stated that 80% of the 51 victims of shootings over a weekend in May 2016 were on the heat list.

Accuracy and how to use the list are not the only issues. Journalists and critics have also raised the issue that the list may reflect racial bias. The heat list is predominately made up of black men. The inputs to the list do not include race or gender or location. But the problem is not explicit bias. The criticism is that the list reflects disproportionate policing in black and poor areas and the fraught relationship between police and young black men. In part the focus of police is in poor areas because there is more crime there. But policing in the US is far from race-blind and racial bias affects law enforcement discretion. The result is more stops by police and greater charge and incarceration rates for non-whites. This leads to an implicit bias in the raw data feeding the algorithm. Biased policing results in a biased algorithm.

Implicit bias may also play a role in programs that predict crime according to place instead of people. Software such as PredPol and HunchLab predict where and when crimes will happen. These approaches use a wide range of data sources to identify crime hot spots and predict when certain crimes are most likely to happen. The products are colour-coded maps showing police the 'what, where and when' of potential crimes.

Predictions based on place face the same problem with bias as those involving people. Police spend more time in certain neighbourhoods and choose to make contact with people who they think are more at risk to break the law or who have already broken the law. This contact results in confrontation and arrests and charges and jail. And these outcomes are the data that feeds the algorithm. It becomes a self-fulfilling prophecy. The software predicts crime in a certain area, more police go to that area and make arrests, and these arrests contribute to the updating of the predictions.

For more on these issues see Andrew Ferguson's book The Rise of Big Data Policing.

Criminal Justice

Algorithms also play a role in contexts downstream from police work. They guide decisions relating to granting bail and parole. Data also plays a growing role in guiding prosecutions. But the example I want to discuss relates to sentencing.

In the US case State v. Loomis the Supreme Court of Wisconsin dealt with the question of whether an algorithmic risk score had a role in sentencing. The algorithm in question was a tool called COMPAS that assesses recidivism risk. Loomis had entered pleas to his charges. But the trial judge rejected the plea deal and gave a harsher sentence in part because Loomis was at high risk to re-offend according to COMPAS. The creator of COMPAS (Northpointe) had refused to disclose the methodology behind the tool to the trial court and the defendant because it was a trade secret. Loomis argued that the use of the tool violated his due process rights because without knowing how COMPAS came up with his risk score he had no basis to challenge its validity.

The court held that the use of an algorithm to inform sentencing did not violate the defendant's due process rights. But it set restrictions on the use of risk scores in sentencing. These were:

1. A court cannot use the COMPAS risk score to decide whether a person goes to jail or for how long.
2. A court must explain the factors independent of the risk score that support the sentence.
3. Any pre-sentence report containing risk scores must set out five cautions with respect to the use of COMPAS, including that the tool may disproportionately classify minority offenders as at a high risk to re-offend.

COMPAS was also the centre of debate outside the courtroom. In 2016 the news outlet ProPublica reviewed data on the use of COMPAS on more than 10,000 people arrested in a county in Florida between 2013 and 2014. In its report it claimed that the tool was biased against blacks. Blacks who did not re-offend were almost twice as likely as whites to be labelled high risk. The contrary was also true. Whites who did re-offend were almost twice as likely as blacks to be labelled low risk.

Northpointe and others challenged the methods and findings of the report. Among other things they argued that COMPAS is not biased because the risk scores do not depend on race. A given risk score gives the same likelihood of re-offending whether the defendant is black or white. But this does not get rid of the false positive problem that blacks were more likely to have a high risk score even though they did not re-offend. Another study of the data pointed out that it is mathematically impossible to get rid of the false positive problem and have parity in risk scores between blacks and whites.

We will have to leave the debates about methods and numbers to the experts. But it's worth talking a little bit more about COMPAS. In 2018 a paper came out showing that COMPAS was no better than untrained humans at predicting recidivism. The research involved 400 volunteers from an online crowdsourcing site. Each of them reviewed the same short description of a defendant that highlighted seven factors (sex, age and 5 factors with respect to crime history). After reading it they responded "yes" or "no" to the question "do you think this person will commit another crime within 2 years?" The paper compared the results with the performance of COMPAS on the same set of defendants. The crowd-based approach was as accurate as COMPAS for both whites and blacks. Adding race to the description did not change the results significantly. The paper also pointed out that statistical predictions using just two (age and past convictions) of the seven factors performed as well as COMPAS that uses up to 137 factors. (See here for a succinct telling of the Loomis and COMPAS story.)

Child Protection

Algorithms and risk scores have also found their way into the world of child welfare and social work. This is not without controversy, to put it mildly. Evoking Minority Report this is what **one critic** had to say: "They are not proposing to rely on psychics in a bathtub. Instead they're proposing something even less reliable: using 'predictive analytics' to decide when to tear apart a family and consign the children to the chaos of foster care."

Those who support data driven tools in child welfare claim that relying on risk scores is more objective than intuitive decisions. They argue that analytics allows social workers to find and focus on children that are at greatest risk. Those who oppose it refer to the same arguments we saw against COMPAS. The algorithms are inaccurate, prone to bias and result in too high a rate of false positives. Good summaries of the debate and the issues are here and here.

A good example of the use of analytics in child protection is the Allegheny Family Screening Tool or the AFST from Pittsburgh. The AFST is in use for call screening at the Pittsburgh hotline for child abuse and neglect. Before the use of the AFST social workers at the hotline had to manually search through large amounts of data and then decide whether to 'screen in' a call and investigate or 'screen out' the call and offer support through the community. There were no protocols to guide the data search or to weigh data points and the search was subject to time pressures and constraints. The AFST on the other hand provides a risk score for the family in question in a matter of moments. The score predicts the likelihood of abuse, placement in foster care or repeat calls to the hotline. A score above a certain level requires review to determine whether an investigation should occur. Full details about the AFST are here.

Like COMPAS and the heat list, the AFST faces criticisms that it carries an implicit bias. According to critics the AFST reflects the disproportionate number of black and poor families caught up in the social welfare system. By relying on biased data the AFST unfairly targets poor families for higher levels of scrutiny. This is a form of "poverty profiling". The County **responded** to this by pointing out that the disproportionality is the result of calls to the hotline (referral bias) and not bias in the screening of those calls (screening bias). The County also noted that poorer families are more likely to be involved with the child welfare system. But it argued that better ways of making child protection decisions should not have to wait for the elimination of poverty.

The development of the AFST was an open process and subject to review by experts and the public. Despite criticism it does seem to have found qualified support. One important factor is ownership. The County owns AFST. This allows a level of public discussion and critique. It is a different matter with COMPAS and most software programs for policing, which are in private hands.

Teacher evaluation

We are stepping outside of public protection for the next example. Here too the algorithm was in private hands and not available to the court or to those who it affects. From 2011 to 2015 the Houston Independent School District used a 'data driven' approach to assess teacher performance. A private software company developed a statistical model to track the impact of a teacher on student test results. The model used changes in students' results on standard tests to create scores to assess how well a teacher was performing. The model resulted in a "Teacher Gain Index" (TGI) that sorted teachers into five levels of performance from well above to well below average. Problems arose when the District began to use the TGI and scores from the model as a basis for firing low performing teachers. It was 'firing by algorithm'. And teachers could not examine the model and methods behind the TGI because they were trade secrets.

The teachers' union **took the District to court**. Among other things it claimed that teachers who were in danger of losing their jobs because of a low score had a right of access to the data and methods to verify the accuracy of the score. On this issue the court sided with the teachers. The decision found that the scores might contain errors for any number of reasons. These included data entry mistakes and computer code glitches. The court stated that, "Algorithms are human creations, and subject to error like any other human endeavour."

The court was also concerned about the inter-relationship of teacher scores. A change in one teacher's score could change the score of all the other teachers. In the court's words, "the accuracy of one score hinges upon accuracy of all." This meant that an error in one score would not be promptly corrected and the results would not be re-run for all teachers. Part of this was due to cost of re-running the data and part of it was about keeping scores stable over time. But the court concluded that a teacher would require access to the data relating to all scores to determine whether his or her score was error-free.

Hospital inspections

The last example is about predicting the risk of poor care in hospitals. In the UK, the Care Quality Commission (CQC) inspects 150 hospital trusts for safety and quality of care. As part of a 'risk-based' approach CQC developed a statistical tool to assess the risk that a trust would provide poor care. The CQC then used the tool to decide which trusts to inspect first or more often. But a **review** found that the attempt to use risk scores to target inspections failed. The review matched the risk scores to the outcome of inspections for the two years after the CQC started to use the tool. The risk scores did not predict inspection outcomes. Nor did the scores more broadly predict which trusts would perform poorly or perform the worst.

The reviewers did suggest some reasons why the tool did not work. One of them was that it was too simple. Some of the indicators that make up the scores may be more important than others but the tool weighted all of them the same. The reviewers also suggested that the tool and the inspectors may simply be measuring different things.

The good, the bad and the ugly of analytics

Our review of these settings is only the tip of the iceberg. But it gives a start to see the good and the bad in data-driven approaches to public protection. We'll start with the bad. The first and obvious point is accuracy. If statistics and scores have an impact on people they need to be reliable and valid. Problems occur because of error or because the methods are unsound or just don't work. You can't use numbers to guide inspection if they don't predict harm. And you don't want to rely on a score with too high a false positive rate.

A related problem is so-called blacklisting. This is the practice of placing people on lists using big data tools. Examples are the heat list or the no-fly list. Putting people on the list due to an error is obviously an issue. But even without error, secret heat or 'threat' lists create issues and face justified scrutiny. It may neither be practical nor prudent to make such lists public as a matter of course. But where an authority takes action against a person on a list there are strong arguments for disclosing to that person information about the list and how he or she got on it.

Secrecy also relates to the disclosure of the nature and methods of an algorithm. There are good reasons for protecting trade secrets. But these reasons should not outweigh due process and the right to challenge a decision that has a material effect on a person's rights and interests. Regulators should compare COMPAS with the AFST in this context. It seems more reasonable for the entity using the software or data to own it rather than a private company.

The issue of bias is a harder problem to deal with than secrecy. Algorithms may not be fair or effective if they reflect existing bias or inequity. There is little use in a risk score that reflects over-policing of a certain group or area, or biased treatment of that group. But there is a growing amount of research on how to detect bias in an algorithm and how to design a fairer one. And questions of bias and fairness must be central to the design, use and review of algorithms in the regulatory context.

Now for the good. Critics spend a lot of time explaining how algorithms are inaccurate or biased. In many respects the criticisms ring true. But we should not overlook human bias. Human decision makers are susceptible to an untold number of biases and mistakes. There are problems with algorithms but in many settings there may be bigger problems with human or intuitive decisions. Research over the past few decades shows that a simple algorithm will likely outperform human judgement if the decision involves prediction. The Nobel Prize winning Daniel Kahneman put it this way in Chapter 21 of his book **Thinking, Fast and Slow**: "The important conclusion from this research is that an algorithm that is constructed on the back of an envelope is often good enough to compete with an optimally weighted formula, and certainly good enough to outdo expert judgement. This logic can be applied in many domains, ranging from the selection of stocks by portfolio managers to the choice of medical treatments by doctors or patients." The real question may not be the shortcomings of algorithms but how much they can **improve on clinical or expert decisions**.

Better predictions should lead to stronger programs. And one of the 'goods' of algorithms is not just about better prediction but how this can lead to better outcomes in the context of the broader setting. A great example is a computer model to identify sources of foodborne illness. New research showed that a model using data from Google sources was more accurate in identifying unsafe restaurants than consumer complaints and routine inspections. Because the model used data in near real time it was quicker to identify an outbreak than traditional methods, which could lead to better public health outcomes. The researchers stated that the best use of the model would be as an adjunct to existing methods, for example as a means to target inspections.

Eight questions to ask about your data program

Our regulator clients are at varying stages with respect to data systems. We'll end with a list of questions to ask before starting or at an early step in a data strategy or project. These questions are also likely relevant to more mature systems. Most of them relate in some way to the issues we saw on our tour of other sectors. And some of them are crucial to avoiding some of the pitfalls of predictive algorithms. Many of these points draw heavily on Andrew Ferguson's advice to police departments in the conclusion to his book **The Rise of Big Data Policing**.

1. Can you identify the harm or the risk that your data program is trying to address? What is the purpose for collecting and analysing the data?
2. Can you defend the data coming into the data program or algorithm? Where are the sources of bias, error or inaccuracy?
3. Can you defend the outputs of the program or algorithm? Do the risk scores or other outputs provide a meaningful measure in the context of reducing harm to the public?
4. Does the use of the data or algorithm have an undue impact on individual autonomy? Are there safeguards against abuse and

a process in place to question or review a score or other output?

5. Is there a regular audit and test of the algorithm? Is there a way to detect bias and reduce it?
6. Is the process of using the algorithm transparent? Can you explain the reason for the algorithm and its design and why it works? Can you explain why it is fair? (This may be more useful than disclosing the actual source code or maths that it is based on.)
7. Who owns it? Should you own it? If you don't own it, what are the risks?
8. Can you get the same results more simply?

This article is a better-organized version of my speaking notes for a seminar at the offices of WeirFoulds LLP on November 14. For those who attended, I have tried to address some of your questions. This topic is new ground for many in the regulated professions sector. I'd welcome hearing from those who are interested in this topic and who wish to pursue it further.

WeirFoulds LLP

www.weirfoulds.com

Toronto Office
4100 – 66 Wellington Street West
PO Box 35, TD Bank Tower
Toronto, ON M5K 1B7

Tel: 416.365.1110
Fax: 416.365.1876

Oakville Office
1320 Cornwall Rd., Suite 201
Oakville, ON L6J 7W5

Tel: 416.365.1110
Fax: 905.829.2035