Matt Herman
May 24, 2018
Infotech – Final Paper

Ethical Considerations for Using Algorithmic Decision Tools in the Public Sector

Recently, there has been a surge of interest in using machine learning algorithms and predictive analytic tools in the public sector. In New York City, algorithms or automated decision tools have been used to assess buildings' fire vulnerability (Reiland 2015), to match 8th graders with appropriate high schools (Tullis 2014), and to identify individuals suspected of a crime using facial recognition (Kanno-Youngs 2018). These applications of machine learning and artificial intelligence may help to increase the efficiency of these city agencies, but at the same time there have been questions raised as to the ethical ramifications and potential for bias that relying on automated decision making may introduce. Moreover, many of these automated decision systems are supplementing or replacing decisions currently made by people such as case workers and judges.

In December 2017, the New York City Council passed legislation that establishes a task force to investigate algorithms used at city agencies (Vacca 2018). In particular, the City Council hopes to "address instances where people are harmed by agency automated decision systems" as well as increase the level of transparency of algorithms that the city uses. The bill is the first of its kind to be passed in the United States and it joins the wider conversation that has been ongoing about algorithmic decision making. In this context, it is essential to consider ethical concerns and potential biases when developing new predictive analytic tools.

There has been long standing debate and conversation in the fields of psychology, medicine, and philosophy as to the decision-making abilities of humans as compared to statistical prediction or algorithmic decision-making. In 1954, the psychologist Paul Meehl published an influential and controversial book, *Clinical vs. Statistical Prediction*, in which he reviewed evidence from many medical studies and argued that on average, "mechanical" decisions (predictions mathematically derived from a combination of inputs) performed better than clinical or human decisions. For Meehl, this was not an indictment of individual clinicians but rather the reality of complex systems and the inability of a human brain to consistently process and weigh a large variety of information. In a later paper on a similar topic, Meehl and colleagues argued that one benefit of mechanical prediction is that given the same inputs, it always leads to the same conclusions. The cite a variety of medical research in which "fatigue, recent experience, or seemingly minor changes in the ordering of information or in the conceptualization of the case or task can produce random fluctuations in judgment" (Dawes, et al. 1989, p10). Additionally, clinicians are susceptible to implicit and explicit biases in their judgements. They may seek out particular information about a case that supports their pre-existing theories while ignoring information that conflicts with it.

Despite a large body of research demonstrating that mechanical, actuarial, or algorithmic prediction is more accurate than clinical or human judgement, there has been a lot of resistance, especially from clinicians, to accepting this premise. Partly, this resistance comes from a legitimate fear of automation and loss of jobs. In radiology, for example, artificial intelligence systems consistently outperform human experts in diagnosis. This has lead to a reevaluation of the role of radiologists in the medical field. George Hinton, a computer science professor and artificial intelligence researcher argued that medical schools "should stop training radiologists now" (quoted in Mukherjee 2017). If jobs in radiology are being replaced by automated decision making tools, it is sensible for radiologists to push back and try to maintain their position in the medical field. But, the resistance to mechanical decision making is more than just a rational fear of unemployment for certain professions, it also comes from the deep seeded belief that human intuition is unique. It even connects to arguments about what makes a human human. These weighty philosophical concepts are underpinning the conversation about algorithmic decision making in many field and that is one of the reasons the debates are so contentious.

Notwithstanding these debates, predictive analytics is widely used in many fields such: advertising (which ads to serve to which people), sports (Sabermetrics), finance (credit scores, high frequency stock trading). But it is only recently this these tools and methods have been used in and by governments. In the same way that many citizens claim to be very concerned about the the way the government is using personal data, yet are willing to share many more intimate details of their lives with large corporations such as Google and Facebook, the public seems to be extremely wary of algorithmic decision making in the public sector, even though it has been going on in the private sector for decades. And there are good reasons for individuals to be more wary when the government is involved; they have the power deprive life and liberty in the way that private corporations do not.

Two of the more controversial applications of algorithmic decision making in the public sector have been in the fields of child services and criminal justice. A recent program created by the Office of Children, Youth and Families (OCYF) in Allegheny County, PA has received a lot of media attention, both positive and negative. Beginning in August 2016, OCYF implemented an algorithmic decision tool called the Allegheny Family Screening Tool. The tool is used when a call reporting child abuse or neglect is received by the county. Upon receiving these reports

OCYF has to decide whether or not to open a formal investigation of the case. There are different legal requirements that may lead to a case being investigated, but prior to 2016 many of these decisions were made by the workers receiving the calls on the basis of the information reported in the call. OCYF hoped to improve this process following a series of high profile failures in which children died after their guardians were previously reported to the county for potential abuse but not investigated by OCYF (Hurley 2018). The county brought in two academic researchers to review five years of reports and outcomes and develop a predictive risk model administrative data that could be used by the county and help staff more accurately screen calls (Chouldechova et al. 2018).

The algorithm uses data taken at the time of the report of abuse, but the primary added value the county hoped it could provide is context for the case. OCYF databases pull data about families including arrest reports, histories of domestic violence and substance abuse, and other risk factors that are associated with child abuse. When a call comes in, the county databases are automatically queried for these variables and a predicted risk score from 1 (lowest risk) to 20 (highest risk) of the particular case is shown to the call screener. In the current iteration of the program, the screener uses this additional information generated by the algorithm along with their personal judgement and other information from the incoming call to determine which cases will be investigated further. Proponents of the system argue that the historical data used in the model contains important predictive information that otherwise would be unavailable to the call screener at the time the decision whether or not to investigate is made. Moreover, if you accept the conclusions of the research reviewed above about clinical and mechanical judgement, the automated screening tool should produce more accurate risk predictions which in turn would lead to more cases that would lead to serious harm to a child being investigated prior to the abuse. After the first 18 months of operation, the algorithm appears to be performing well (Hurley 2018). The percentage of low-risk cases that are recommended for investigation has decreased from fifty percent to close to just above thirty four percent. This means that there is more time and resources available for staff to investigate and prevent the higher risk cases.

Even though there have been preliminary successes, the Allegheny County example is not without its dissenters. In particular, Virginia Eubanks dedicates one chapter of her book, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (2018), to describing the problematic aspects of the Allegheny Family Screening Tool. The core source of Eubanks' criticism is that racial minorities are overrepresented in the data that was used to generate the original model as well as the new data that is fed into it to predict each new case. In particular, much of the data used to to predict the risk scores comes information about the use of public programs such as Medicaid, food stamps, drug treatment as more. Inherently, this means that individuals who use these programs are more likely to be picked up by the screening tool, while families who choose not to use county services are less likely to be scored highly. Eubanks notes that "a quarter of the variables that the AFST uses to predict abuse and neglect are direct measures of poverty….Another quarter measure interaction with juvenile probation and CYF itself, systems that are disproportionately focused on poor and working-class communities, especially communities of color" (Eubanks 2018 p146). This explicit relationship of the volume of predictor variables to poverty and racial inequality lead Eubanks to conclude the the screening tool largely serves to report have many public resources a family has used, rather than predicting the risk of harm to their children.

An even more controversial application of algorithmic prediction tools in the public sector is in the criminal justice field. Increasingly, courts are using algorithms to determine sentencing guidelines for those convicted of a crime. In addition to weighing the severity of a crime and the defendant's prior history, Judges look to mechanized assessment tools to determine the risk of recidivism for individual cases. A leading assessment tool is COMPAS, the Correctional Offender Management Profiling for Alternative Sanctions, developed by Equivant, a private company. A 137-item questionnaire is administered to defendants and the company's proprietary algorithm spits out a risk score. COMPAS was the subject of a Wisconsin Supreme Court case in which a defendant who was sentenced to six years in prison contended that the punishment was unconstitutional because he and his lawyers were unable to review the process by which the algorithm rated him a high risk for recidivism. Even though the tool is used by the state, because it used a proprietary algorithm, it was not publically available (Yong 2018).

But the secretive nature of the algorithm is not the only issue with COMPAS. In a study published in 2018, researchers from Dartmouth College tested the quality of the its predictions. The recruited 400 volunteers with no special criminal justice training and gave them summaries of crimes committed and the background of the defendant. They asked each participant to assess if each defendant would commit another crime. Next, they compared the individual predictions to the predictions made by COMPAS and found no significant difference in the predictions made by the humans and the predictions made by the algorithm; both predicted about 65% of cases correctly (Dressel and Farid 2018). Beyond the level of bias or fairness of using a proprietary algorithm, this finding suggests the COMPAS algorithm may simply not be working well. This finding goes against the research discussed above showing that mechanical prediction is almost always more accurate than clinical judgment. So it is possible that one problem with COMPAS is in its modeling and assumptions. But because it is unable to be inspected by researchers, it is impossible to determine if the algorithm has to the potential to be improved so that it's predictions exceed the ability of humans.

Both COMPAS and the Allegheny Family Screening Tool have caused many questions to be asked about the role of algorithmic decision making in the public sector. One reason these two cases of gotten a lot of attention is that they both represent serious incursions into private citizens lives by the government. A judge has the power to put someone in prison for the rest of their life. The Office of Children, Youth, and Families can remove a child from her birth parents and put her in foster care. There are serious and obvious negative human costs associated with these decisions and so extra care is warranted in the use of automated decision tools. On the other hand, if a city were developing an algorithm to select the winner in a fun, "greenest block of the year" contest, the consequences would be relatively minor and less likely to cause controversy or backlash. In cases when an individual's life or liberty is at risk from a decision made by the state, a higher level of accountability and transparency is warranted. So an important consideration for the government is the extent to which algorithmic decisions will cause punitive or negative harm. In other cases, the goal of a predictive risk model may be to decide which families are eligible to receive additional city services or interventions. In these cases, there is less potential to harm to the individual and it may be more appropriate to use automated decision making tools.

Another concern is that certain algorithms, even if they aren't property of a private company, are very hard to interpret and understand. Many new machine learning techniques involve extreme non-linearity and other "black-box" processes that limit their interpretability. In contrast to more traditional regression-type analyses, modern algorithms are unable to explain how specific variables influence the output of the model. It is a case wherein the tools are designed the maximize their predictive value and the expense of explanatory ability. And this lack of explanation has the potential to represent an unfair implementation of an automated decision-making tool. It is then, the challenge of these methods to improve the predictive power of existing tools to a great enough extent that the shortcomings in interpretability are outweighed by the increased accuracy of the predictions.

Overall, there are many challenges and opportunities for predictive analytics in the public sector. As Virginia Eubanks argues, there is the risk of the reproduction of systemic biases because these algorithms are trained on existing data that has been collected and generated within a biased system. In a world where black people are more likely to be arrested than white people, even for the same crime, using arrest data to predict recidivism might "teach" the model that black people are more likely to *commit* a crime, even though it only shows that they are more likely to be *arrested*. But it is important to consider the status quo when assessing the value and fairness of algorithms. The current system relies on individual call center workers or police officers or judges to predict the outcomes of people. The same systemic biases that exist in administrative data also exist in people. People are key parts of the system that have led to the current inequalities represented in the data. So the question should not be: are algorithms transparent and fair, but rather are algorithms *more* transparent and fair than leaving these predictions to humans?

Human decision making is certainly flawed, but may be flawed in different ways than algorithmic decision making. One possibility to balance the competing concerns between human and mechanical judgement is to create a system that can integrate both machine prediction and expert human opinion. The Allegheny County system does a version of this whereby the call screeners initially rate the severity of the risk of the call and then the automated screening tool gives its “opinion.” Individuals are able to override a certain portion of the decisions the algorithm makes and the algorithm may push an individual to consider additional or different contextual evidence when making their decision.

Finally, a key aspect of using machine prediction in the public section is the process by which these tools are built. The process of generating models and algorithms should involve many stakeholders from government and outside organizations (Shroff 2017). This collaborative processes emphasizes communication between technical and policy staff as well as end users. In this way, the implications and applications of algorithmic decision making can be taken into account throughout the development process. Another important recommendation is for algorithm developers to estimate the impact of the tool on vulnerable populations, especially racial minorities. Continued evaluation after implementation should also monitor the extent to which the algorithm has differential outcomes for minorities (Shroff 2017).

References

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. “A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions.” Pp. 134–48 in *Conference on Fairness, Accountability and Transparency*.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. “Clinical Versus Actuarial Judgement.” *Science; Washington* 243(4899):1668.

Dressel, Julia and Hany Farid. 2018. “The Accuracy, Fairness, and Limits of Predicting Recidivism.” *Science Advances* 4(1):eaao5580.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press.

Hurley, Dan. 2018. “Can an Algorithm Tell When Kids Are in Danger?” *The New York Times*, January 2.

Kanno-Youngs, Zolan. 2018. “Facial Recognition Could Move Beyond Mug Shots - WSJ.” *Wall Street Journal*, April 4.

Meehl, Paul E. 1996 (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

Mukherjee, Siddhartha. 2017. “A.I. Versus M.D.” *The New Yorker*, March 27.

Reiland, Randy. 2015. “How Data and a Good Algorithm Can Help Predict Where Fires Will Start.” *Smithsonian Mag*.

Shroff, Ravi. 2017. “Predictive Analytics for City Agencies: Lessons from Children’s Services.” *Big Data* 5(3):189–96.

Tullis, Tracy. 2014. “How Game Theory Helped Improve New York City’s High School Application Process.” *New York Times*, December 5.

Vacca, James. 2018. *Automated Decision Systems Used by Agencies.* New York City Council.

Yong, Ed. 2018. “A Popular Algorithm Is No Better at Predicting Crimes Than Random People.” *The Atlantic*, January 17.