University of Notre Dame

Algorithmic Bias in the Criminal Justice System

SOC 10033: Introduction to Social Problems

Dieter Erben Vasconcelos
12-11-2017

Introduction

The idea of an algorithm has been around for hundreds of years, as Egyptians and Babylonians developed algorithms to multiply numbers, find square roots or for factorization. Yet, the idea of an algorithm as we know it today became popular in 1928 with David Hilbert's "Entscheidungsproblem" (decision problem). This was based on the concept of a problem with a set of inputs and a yes or no output. The algorithm would receive an input and give back a 0 or a 1 (a no or a yes). However, during the 21st century, we have become overly reliant on modern algorithms based on data to make decisions for us. Algorithms are not just giving us a yes or a no, but making complex decisions and dictating what we see online, how we behave and how we make decisions. Collecting data is not the problem anymore, since according to IBM, about 90% of all the data in the world has been generated in the past few years (Wall 2014). The current problem is what to do with all this data, and as this is a new territory, sometimes we forget the consequences of these advancements in the field. Predictive analytics has become very important in many fields such as marketing, retail, healthcare, insurance, financial services like credit scoring, and more. By using many statistical techniques and big data, we try to analyze historical and current behaviors to predict the future. Yet, the model is only as good as the data allows it to be, based on the people and processes that collected the data (O'Neil 2016). If our original data is biased, our model will be biased as well. Many data scientists focus on their specific definition of success based on the context they are working on, not considering how the deployment of this algorithms could create unintended consequences. Relying on models without considering the possible errors can be catastrophic, as I will mention on this paper. I will focus on certain applications related to the criminal justice system and surveillance.

Source

The criminal justice system has some history regarding risk assessment. In the 1980s, a movement called "selective incapacitation" was based on the idea that a set of people should be identified based on their proneness to violence, and give them longer sentences to reduce crime rate (Mathiesen 1998). The likelihood of re-offending plays a big role here, and statistical models were used to determine this. The focus of this movement was to decrease crime rates by incarcerating the most dangerous criminals and reduce mass incarceration as well. However, the main problem with this movement was the number of false-positives the predictions were given, meaning incorrectly labeling people that should not have been labeled. Another problem was the assumption that career criminals, or habitual offenders, can be easily identified based on criminal history and other personal characteristics. Yet, a study done by the RAND Corporation showed that while the models were relatively accurate (76%) at predicting low-rate offenders, they were very inaccurate with high-rate offenders (45%), resulting with a 55% rate of false-positives (Cohen 1983). This was significant because it violated the presumption of innocence, leading to the use of predictive algorithms in other areas of criminal justice such as the risk of not showing up to court or the risk of performing another crime before the trial. But when it comes to sentencing, there are many other variables to consider.

As the Model Penal Code says, the severity of the punishment and its length should be related to various theories of punishment such as "individual retribution, rehabilitation, deterrence and incapacitation" (Kehl and Kessler 2017). The problem with a sentence is that it is not a binary decision of trying to decide if someone should go to jail or not. The punishment can change a lot based on the severity and duration of the sentence. Basing it on a consequentialist approach to punishment, according to Sonja Starr, there is not a lot of evidence showing that a

longer sentence affects the probability of a criminal to commit future crimes (Starr 2014). Besides, punishment is not solely based on the consequentialist theory. Effects like deterrence and the retributive theory of punishment are hard to quantify and to include in an algorithm. A letter from the Department of Justice stated: "experience and analysis of current risk assessment tools demonstrate that utilizing such tools for determining prison sentences to be served will have a disparate and adverse impact on offenders from poor communities already struggling with many social ills. The touchstone of our justice system is equal justice, and we think sentences based excessively on risk assessment instruments will likely undermine this principle" (Holder, Department of Justice 2014).

Related to Holder's argument, the use of data analytics and algorithms in sentencing can be risky due to the inclusion of factors that are not controlled by the defendant. While previous crime data is part of the current models used, it is not enough to give an accurate prediction of crime. It becomes a problem when they start to look at a defendant's economic status, education level, demographics and more. First, a lot of these factors are not under the control of the defendant. They are tied with many social problems that are part of society right now, and giving them even more value in an algorithm is just worsening these problems. These data would show the chances of a minority offender to commit a crime higher than that of a white-collar criminal (Gerstein 2014). Other problems with the use of algorithms in sentencing arose with the case of *State v. Loomis*, where the legality of using risk-assessment software in criminal sentencing was studied, since it could be a violation of due process. Eric Loomis, the defendant, argued that COMPAS, the company that created the software, was not transparent with their software so there was no way to test the accuracy of the risk score (Palazzolo 2016). Besides, it was not an individualized sentence since they were using "characteristics of a larger group to make an

inference about his personal likelihood to commit future crimes" (Kehl and Kessley 2017). This relates to the ecological fallacy, where the characteristics of a larger group are assumed to be true for a specific member of this group. While the case ended up going against Loomis, citing that the algorithm score was just one of the factors in the decision, it gave the court a chance to define how they would handle future cases related to this one. It showed us that, since these are new circumstances, there are no precedents to refer to and these cases will just become more common in the future. We can see another example of the ecological fallacy in a court decision for the Washington gubernatorial election of 2004, where they were trying to figure out the identity of invalid voters based on the precincts where these votes were casted. In other words, they argued that the invalid votes should be assigned based on the voting patterns of the precinct where the vote was casted. A report by a professor of the University of Washington pointed out this mistake, arguing that they could not assume that the characteristics of the average applied to the individual. The missing votes could have been unrepresentative of the precinct's voting trends. (Adolph 2005). Besides, using data pertaining to a larger group goes against the penal code, since the sentence must be individual.

There are some sociological theories that explain in a more theoretical way the possible causes and consequences of this problem. Marion Fourcade argues that there are three types of classificatory judgments: nominal, cardinal and ordinal. Nominal judgments are based on essence, or on what the subject is. This type of judgment first started in 1735 with Carl Linnaeus' *System Naturae* where he applied this classification to plants and animals. As more information is acquired regarding new specimens, the different categories are updated based on similarities and differences (Fourcade 2016). On our specific case regarding criminal justice, if we are making nominal judgments, we must consider homophily and embeddedness. An individual's

personal bio is created by the ways that it is connected to other people, groups, processes or places, and spatial embeddedness can be a factor too when creating a training database for a policing model. Operationalization is another problem seen. By using previous crime data (since it is what is available) we could be predicting future policing instead of predicting crime. The second classificatory judgment, cardinal, also plays a role in this example (Fourcade 2016). By aggregating numerical values, we could fall into the trap of an ecological fallacy or an anthropomorphic fallacy, making conclusions about individuals based on the groups that individuals belong to, or treating a group as if they were one person. The last classificatory judgment, ordinal, is based on ranking of different groups. In contrast with nominal judgment, ordinal judgment implies that there is a different valuation of the different groups, with one being higher and the other one lower (Fourcade 2016). These rankings can be very subjective, as we are talking about human beings and ranking characteristics of who we are or what we do. Ranking cities/counties/regions based on previous crime activity can lead to problems related to aggregates. We want all to align perfectly in these aggregates, yet this is not how it works. Not all groups are mutually exclusive and collectively exhaustive. As I move on to discuss the consequences of algorithmic bias, I will start with a simple example to provide some context and continue with a more complex example related to the justice system.

Consequences

An example that describes this problem is discrimination with online ads. Based on a study done by Latanya Sweeney of Harvard, Google Ads generate ads suggestive of an arrest based on the name of the person on the search term. A "black-identifying name was 25% more likely to get an ad suggestive of an arrest record" (Sweeney 2013). This doesn't mean that Google's ads are intentionally biased. Since their algorithms are based on previous data, it is instead showing us how racially biased our society is. Yet, there are more serious examples as the one of predictive policing software.

Starting in 2008 at the Los Angeles Police Department (Perry 2013), predictive policing programs try to predict certain crimes in a similar way to how scientists predict earthquake aftershocks. Currently, many states such as California, Arizona, Illinois, Tennessee and more are using predictive policing programs in their police departments (Friend 2013). This technology was so groundbreaking that TIME magazine called it one of "The 50 Best Inventions" of 2011 (Grossman, Thompson, Kluger 2011). Yet, the problem with these models is that they are trained based on biased data. This software uses previous data to learn and try to reproduce patterns. The results can be an ineffective model, or even worse, a discriminatory model (Lum and Isaac 2016). Police data may not include all criminal offenses, or be sampled randomly. Since this data is based on what previous officers have recorded, there is always the chance for error or bias. More surveillance leads to more recorded crimes, and an algorithm could be trained to predict crime based on recorded crimes, which are not the same as actual crimes. Therefore, neighborhoods with high surveillance would have a higher percentage of their crimes recorded compared to other neighborhoods (Theory of Deterrence). This problem is made even worse when we consider how minority ethnic communities have a history of being treated differently

by the police, targeted with specific forms of policing (Bowling 2003). “Empirical evidence suggests that police officers – either implicitly or explicitly – consider race and ethnicity in their determination of which persons to detain and search and which neighbourhoods to patrol” (Lum and Isaac 2016). An analysis of COMPAS, a predictive policing tool, showed that “black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk” (Mattu 2016). If more data is being recorded on certain neighborhoods that the police usually patrols, these would be over represented in the database. Yet, police bias is not the only way that this model is affected. Many crimes are reported by citizens, making the source of this bias “community-driven rather than police-driven” (Lum and Isaac 2016). Over-policing certain communities have shown an increase in mental and physical health problems, as well as more opportunities for police violence.

Algorithmic bias can have a compounding impact when used in the justice system, and we can see some type of echo-chamber effect. If the data collected is not objective and already includes bias, then the models will be trained with this bias. (Aruna Kumari 2015). Since a model is only as good as the data that gets fed to it, then it would make its predictions biased. People are being entered into models and algorithms only as data points, without considering the veracity of the data. A vicious cycle can be seen where the prediction algorithm would be as biased (and even considered racists) as the data used to build it. The quality of analysis can be damaged if all data points are treated as the same (Aruna Kumari 2015).

Lastly, in the process of predictive policing, there is a difference between predicting crime hot spots and predicting offenders, using predictive maps leading to surveillance or network analysis leading to home visits (Hvistendahl 2017). Even if we assume that a model is

right most of the time, there can be a debate between surveillance efficiency/security and freedom. For example, crime rates could be reduced drastically if everyone had a curfew and had to stay in their house for most of the day. But is it worth it to trade security for freedom? This also relates to the culture of each country, but at least in the United States, freedom is held in high regard, and increasing surveillance decreases freedom. As Franks states, "There is general agreement across the political spectrum that surveillance is harmful at least in part because it constrains individual autonomy and expression, which in turn jeopardizes the possibility of a truly democratic society" (Franks 2017).

Solution

As data becomes more openly available and widely used, we must be keep in mind the concept of information justice. We tend to believe that the availability of data to everyone is a positive, as everyone can access it and use it for their own benefit, democratizing information and giving control to the people (Johnson 2014). However, Johnson argues that there are three main problems with the open data movement, all related to the failure to understand how data was constructed. I will focus on the first problem: "the embedding of social privilege in datasets as the data is constructed" (Johnson 2014). We need to recognize that "whether by design or as unintended consequences, the process of constructing data builds social values and patterns of privilege into the data" (Johnson 2014, p.265). The justice (or lack of) is a characteristic of the data collected, as the data is describing our society. A lot of social data is collected when an individual interacts with an organization like a business or the government. However, the interaction between these groups differs between people. Additionally, the collection of data can cause some of these problems to be overrepresented or underrepresented. An example of this can be seen with census data, where minorities can be undercounted due to many factors such as address errors, lack of civic responsibility and distrust of the government (Prewitt 2010). While this is not a problem that can be easily addressed and solved, the first step should be the acknowledgment of the problem. When building models, algorithms and predictions, there is a need to keep in mind these problems and recognize that the model will include the real nature of our society, with all its injustices and biases. Data architects decide what information to collect, what information is missing, and what information to use when building models, so it is crucial to consider how the choices made here will affect the algorithms.

Fortunately, there are organizations that are working to solve some of these issues. There are organizations such as The Algorithmic Justice League or Algorithm Watch that have the goal of evaluating and identifying bias in current algorithms (Pelzel 2017). The Algorithmic Justice League looks to raise awareness about algorithmic bias, providing a space to report experiences with coded bias and develop practices for accountability during the design and development of systems. The AJL has staff from coders and academics to companies, activists and legislators involved in the fight against algorithmic bias. Joy Buolamwini, the founder of AJL, argues that part of the problem is a lack of diversity, both in the data collected and in the tech industry itself. "If you test your system on people who look like you and it works fine then you're never going to know that there's a problem" Joy says (Kleinman 2017).

Suresh Venkatasubramanian, a professor at the University of Utah, proposes three solutions to this problem. The first one is the creation of better and more diverse data sets when training models. The data should represent the population that will be affected by the algorithm, so the training set should be representative of both the testing set and the real set. Second, the best practices should be shared among software vendors. Many of these problems are relatively new, so more will continue to arise, and these need to be shared so they can be tackled together. Lastly, the code and data behind algorithms and how they make decisions should be transparent, so that biases can be understood and considered when employing algorithms.

The Algorithm Watch is another non-profit trying to show the social relevance on algorithmic decision-making processes. Their ADM (algorithmic decision making) manifesto states that (ADM Manifesto):

1. ADM is never neutral
2. The creator of ADM is responsible for the results it generates

3. ADMs have to be comprehensible so they can be held accountable
4. Democratic societies need to achieve intelligibility of ADM with of technologies, regulation, and oversight institutions.
5. We must decide how much of our freedom we allow ADM to preempt.

While the ADM is still not widely known, it is at least an attempt to control the decisions made by algorithms and consider multiple factors before these are made. As we continue to employ algorithms and make mistakes, we need to consider the effects of these errors and how to try to avoid them in future cases.

Discussion and Future Research

The rate at which technology is advancing is both exciting and alarming. Often, we don't take the time to consider if we should go ahead with a technological advance. The decision is based on our capabilities, not on what should be done (can we vs. should we). We look to optimize decisions by increasing efficiency, and since computers can help us do this, we tend to forget the effects of this decisions. Algorithms can help us make decisions faster, and can analyze data in a way that humans are not capable of. The volume and velocity that they can handle is far superior to what we can do. However, this doesn't mean that we should trust these decisions blindly. A model is only as good as the data it receives. As a common saying in data science says, "garbage in, garbage out". If the data we use to build models is not accurate, is biased or even racist, then the predictions coming from the model will be like the data.

This problem is relatively new, yet it will become more important and will affect our society even more in the future. While some of the sources and consequences are available, the solutions are significantly behind. Our society is still trying to understand the massive effects that an algorithm can have, and there have been no signs to show that we will stop using them to make decisions. On the contrary: more and more decisions are automated now, and will continue to be like this in the future. This research paper is limited to the recent history of the problem, yet as more scientists and researchers focus on the problem, more causes and consequences will be found, and hopefully solutions as well. While the power of these algorithmic models is undeniable, we must understand the responsibility that comes with these advancements in technology. We can't let technology decide for us.

References

Adolph, C. (2005). *Report on the 2004 Washington Gubernatorial Election* (Rep.).

Aruna Kumari, D., Tejeswani, N., Sravani, G., & Phani Krishna, R. (2015). Echo Chamber Effect In Big Data. International Journal of Computer Science and Mobile Computing, 4(4), 476-479. Retrieved from http://ijcsmc.com/docs/papers/April2015/V4I4201599a10.pdf

Bowling, Ben and Phillips, Coretta (2003). *Policing ethnic minority communities.* In: Newburn, Tim, (ed.) Handbook of Policing. Willan Publishing, Devon, UK, pp. 528-555. ISBN 9781843920199

Cohen, Jacqueline. "Incapacitation as a strategy for crime control: Possibilities and pitfalls." *Crime and justice* 5 (1983): 48-49.

Department of Justice Letter by Eric Holder: "The Promise and Danger of Data Analytics in Sentencing and Corrections Policy". (2014).

Deterrence, Theory of. (n.d.). *The Social History of Crime and Punishment in America: An Encylopedia.* doi:10.4135/9781452218427.n184

Fourcade, M. (2016). Ordinalization: Lewis A. Coser Memorial Award for Theoretical Agenda Setting 2014. *Sociological Theory*, *34*(3), 175-195.

Fourcade, M., & Healy, K. (2016). Seeing like a market. *Socio-Economic Review*, *15*(1), 9-29.

Franks, M. A. (2017). Democratic Surveillance. *Harvard Journal of Law & Technology*, *30*(2). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2863343

Friend, Z. (2013, January 22). Predictive Policing: Using Technology to Reduce Crime. Retrieved September 25, 2017, from https://leb.fbi.gov/2013/april/predictive-policing-using-technology-to-reduce-crime

Gerstein, Josh. “Holder: No 'Moneyball' in sentencing.” *POLITICO*, 1 Aug. 2014, www.politico.com/blogs/under-the-radar/2014/08/holder-no-moneyball-in-sentencing-193262.

Grossman, L., Thompson, M., & Kluger, J. (2011, November 28). The 50 Best Inventions. Retrieved September 25, 2017, from http://content.time.com/time/magazine/article/0,9171,2099708-13,00.html

Hvistendahl, M. (2017, July 26). Can 'predictive policing' prevent crime before it happens? Retrieved September 25, 2017, from http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

Johnson, J. A. (2014). From Open Data to Information Justice. *Ethics and Information Technology,* 263-274. doi:10.2139/ssrn.2241092

Kehl, Danielle Leah, and Sam Ari Kessler. "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing." (2017).

Kleinman, Z. (2017, April 14). Artificial intelligence: How to avoid racist algorithms. Retrieved December 09, 2017, from http://www.bbc.com/news/technology-39533308

Lamb, Evelyn. “Review: Weapons of Math Destruction.” *Scientific American Blog Network*, 31 Aug. 2016, www.blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14-19.

Mathiesen, Thomas. “Selective Incapacitation Revisited.” *Law and Human Behavior*, vol. 22, no. 4. (1998): 455–469.

Mattu, J. L. (2016, May). How We Analyzed the COMPAS Recidivism Algorithm. Retrieved September 25, 2017, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?lipi=urn%3Ali%3Apage%3Ad_flagship3_pulse_read%3BG6qM%2Fg%2BhRBOYKDVFh5nwrg%3D%3D

O'Neil, C. (2016, February 04). How to Bring Better Ethics to Data Science. Retrieved September 24, 2017, from http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html

Palazzolo, Joe. “Wisconsin Supreme Court to Rule on Predictive Algorithms Used in Sentencing.” *The Wall Street Journal*, 5 June 2016, www.wsj.com/articles/wisconsin-supreme-court-to-rule-on-predictive-algorithms-used-in-sentencing-1465119008

Pelzel, F. (2017, August 28). Big Data will be biased, if we let it. Retrieved September 25, 2017, from https://www.linkedin.com/pulse/big-data-racist-we-let-federica-pelzel/

Prewitt, K. (2010). The U.S. decennial census: Politics and political science. *Annual Review of Political Science*, 13(1), 237–254.

Starr, Sonja B. "Evidence-based sentencing and the scientific rationalization of discrimination." *Stan. L. Rev.* 66 (2014): 803, 805-56.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 10.

Wall, M. (2014, March 04). Big Data: Are you ready for blast-off? Retrieved September 24, 2017, from http://www.bbc.com/news/business-26383058

Walter L. Perry (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation. p. 4.

Wortley, R. K., & Townsley, M. (Eds.). (2016). *Environmental criminology and crime analysis* (Vol. 18).

"The ADM Manifesto." AlgorithmWatch, www.algorithmwatch.org/en/the-adm-manifesto/

"Algorithmic Justice League Mission" Algorithmic Justice League, https://www.ajlunited.org/home