A Machine Learning system beyond the code:

the case of the COMPAS algorithm

Author: Catalina Alzate

School of Arts, Technology and Emerging Communication

University of Texas at Dallas

The COMPAS algorithm developed by the company Northpointe is an online tool used for assessing a person's risk of recidivism in several courts in the U.S. This tool was analyzed by an independent news agency and labeled as biased against black people. Since this 2016 expose, journalists have extensively cited this as an emblematic example for showing how Machine Learning systems can have a negative effect in society. However, the news articles focus in describing the technical aspects of this technology without placing accountability for the racist bias and ignoring the larger sociopolitical context where this technology operates. The techno-centric accounts situate the technology as something accessible only to scientists and programmers and therefore, as a reader not familiarized with Machine Learning systems, it is difficult to envision how the problem of bias can be solved. In this paper I analyze the narratives about the COMPAS algorithm from five online news articles, complement their perspectives by situating factors in larger contexts, tracing connections amongst them and contributing with a set of possible scenarios for change. Through this process, I try to demystify the notion that a Machine Learning system is unattainable for users and only transformable by programmers, and bring it closer to the audience by exposing how it needs a number of actors and contexts to operate.

Introduction to the COMPAS algorithm

When a person is arrested and taken into court, the judge will determine his or her probability of committing a crime in the future, by asking a series of questions during the hearing and input the responses into COMPAS, a survey-like document, available on a web browser. Some of these questions will be already filled by the software, since the ID of the person will indicate criminal records and some "static factors" such as age at first arrest and gender (Practitioner, 5). The rest of the questions are entered manually. When the survey is complete, the software analyses the information and provides a risk score as a number from 1 to 10 according to the likelihood of the person of committing a new crime over the next two years. According to this number, the judge should assign a rehabilitation process for the offender. The COMPAS algorithm serves the function of predicting risk and helping judges decide on rehabilitation needs (Practitioner, 8).

ProPublica, an investigative news organization, studied the data of people from the state of Florida who were scored used the COMPAS system. While race is not an explicit question in the initial survey, ProPublica exposed how black people who are sentenced and do not reoffend, are twice as likely than white people to be scored as high risk (Angwin et al.) The title of this study is "Machine Bias" and in a short period of time, it became a popular reference in news articles that highlighted how this algorithm is skewed. I have chosen five sources of online news articles by The New York Times, The Wall Street Journal, The Washington Post, The Atlantic and The Guardian and identified four common themes across them. In the following section, I introduce and complement these themes using supporting arguments from academic sources, and describe how the narratives in the news stagnate the conversation about the future of this technology. In contrast, I transcend techno-centric discussions that focus on the COMPAS algorithm, by including the company that created the technology, the criminal justice system and the machine

learning community as active stakeholders that act in tandem to allow this technology to function and express racial biases.

1. Advocating for transparency is insufficient

The first common theme across the news articles is the need of transparency to fully understand how the algorithm generates biased results. For example, the Washington Post affirms: "Northpointe has refused to disclose the details of its proprietary algorithm, making it impossible to fully assess the extent to which it may be unfair, however inadvertently" (Corbett et al.). This assertion is echoed and cited in the other news articles, suggesting that accessing the code would contribute to fix the problem. Mike Ananny and Kate Crawford, academic researchers in the Machine Learning field, expose how the idea of transparency is limited in two fronts: on one hand, the code that constitutes COMPAS, as any other Machine Learning system, is a complex mathematical network that increases in size and density by computer operations alone. Given its complexity, it would be hard to decipher it and there may not competent audiences that comprehend it (982). Whereas this holds true for many algorithms, complexity leaves us with no hope for change. In contrast, Safiya Umoja in her book *Algorithms of oppression*, has demonstrated how Google, after denying intent or responsibility to harm, has actually manipulated the algorithms when accused of racist biases (82). By facing these two views, I argue that advocating for transparency is not enough. The code may not be necessarily made public, but it has to be revisited by the programmers in Northpointe and this can be achieved if there is enough pressure and public demand. Transparency evokes the idea of visibility of the code, and we must include the people in charge of such visibility in the conversation, as a first step towards accountability.

The second limitation pointed by Annany and Crawford, is that transparency refers to a specific algorithm in isolation. However, "there is no single system to *see inside* when the system

itself is distributed among and embedded within environments that define its operation" (982). In other words, the COMPAS algorithm is more than code. It is a set of relations between the code and its environment. As we will see in the following section, the issue of bias is not exclusive of the algorithm but very much part of its users and enablers.

2. The root biases of the algorithm exist within the criminal justice system

The news articles mention how the training data for the COMPAS algorithm mirror our biases in society. For example, the article by The New York Times mentions: "An algorithm that accurately reflects our world also necessarily reflects our biases" (Israni). The claims related to the training data in this and other news articles, dilute the responsibility to society in general. It is necessary to contextualize these claims to understand the origins of the bias. The most relevant training data used for the COMPAS algorithm are criminal records, which are generated within the criminal justice system and it is indeed this system that ends up being mirrored in the operations of the algorithm. As Katheryn Russel-Brown explains, race has historically mattered within the criminal justice system in many explicit and implicit ways:

> When compared with Whites, people of color, particularly African Americans, Latinos, and Native Americans, experience racially disparate treatment at every stage, from arrest to incarceration. These outcomes are sometimes the result of intentional racial discrimination. However, we now know that these disparities reflect more than conscious racial animus. Implicit racial bias is another road that leads to racial inequities in policing practices, court processes, and correctional treatment. (Russell-Brown, 191)

Russel is referring here to the biases that law enforcement agents express when they are in direct contact with people of color. Interestingly, the three processes that she mentions: policing practices, court processes and correctional treatment, are all recorded as databases, and they

constitute the criminal record of a person. This is the most important reference to decide if a person will re-offend. Consequently, if the practices of policing and sentencing are already biased against black people, this bias will be encoded in the record of such practices and become the training data for an algorithm like COMPAS.

Additionally, there are other datasets included in COMPAS. Tafari Mbadiwe explains that even if race is not made explicit in such datasets, they can be *race proxy*, which means that "while it doesn't explicitly *refer* to race, it does *correlate* with race - examples include ZIP code, socioeconomic status, and previous arrests. Typically, risk- assessment systems like COMPAS do not use race-related data but do incorporate proxies." (19). If the databases related to crime already exhibit biases, anyone who tries to use them to design an algorithm will get skewed results. This again leaves us with no hope for the creation of better algorithms. However, we must recognize that datasets are *chosen* to be part of a machine learning system. In other words, there is an underlying system of classification and sorting of information that has been designed as the basis of this algorithm. By influencing the choice of databases to use, we could imagine how a recidivism algorithm could produce different results. Nevertheless, deciding about the training data in this context cannot be a responsibility of computer programmers. These choices correspond to larger questions about fairness and crime reduction, and the criminal justice system must answer those.

3. A simpler version of COMPAS will also exhibit bias

The narratives by The Atlantic and The Guardian mostly focus on questioning the usefulness and complexity of the COMPAS algorithm. This is informed by the research published by Julia Dressel and Hany Farid, who crowdsourced 700 respondents through Amazon Turk, provided them with seven variables from records of offenders in the state of Florida, and requested them to predict if the person would reoffend or not in the next two years. According to their results, guesses from

random respondents exhibit the same accuracy than the COMPAS algorithm (Yong). After analyzing this work, the news articles question one of the premises that this algorithm is built upon: that a machine will generate a better judgment than a human being.

A second study conducted by Dressel and Farid involved the design of a predictive algorithm that works only with two risk factors: age and number of previous convictions. After being tested, it resulted again in a very similar accuracy percentage to COMPAS, therefore questioning if its complexity is even necessary. These studies do not necessarily suggest that COMPAS is useless, but rather demonstrate that algorithms must be understood and tested "before hinging people´s lives on it" (qtd. in Yong). The missing part of the argument is however, that even if the algorithm uses only two factors: race and previous conditions, the results will also be biased, since black young men are disproportionally incarcerated in the U.S. This is therefore not a matter of complexity of the technology, but rather about the overall parameters that decide its functioning.

4. Mathematical inevitabilities should not compromise social justice

The Washington Post reanalyzed the data collected by ProPublica and exposed a discrepancy between Northpointe and ProPublica's definition of fairness. The claim by ProPublica is that black people who are sentenced and do not reoffend, are twice as likely than white people to be scored as "high risk" by the algorithm (Angwin et al.), and the response of Northpointe is that among those who are labeled "high-risk", black defendants and white defendants will go on to re-offend at the same rates (Mbadiwe, 12). Both of these scenarios are valid but they depart from two different assumptions: Northpointe claims that the algorithm is fair because it is equally precise regardless of race, and ProPublica argues that for the algorithm to be fair, it should make mistakes at the same rate regardless of race.

The statistical analysis by The Washington Post demonstrates that each definition of fairness tells a different story, always disfavoring one group over another. They label this as a mathematical inevitability: "The imbalance ProPublica highlighted will always occur" (Corbett et al). According to this logic, a new definition can always be crafted in order to justify the technology. However, the bottom line is that a private company should not define what fairness mean for criminal justice. The legal system should first clarify the questions that are at stake. For example: Should prison programming focus on inmates at high risk of recidivism?, should risk assessment be incorporated into sentencing?, should there be a decreased focus on long prison sentences? (James, 3)

Overall limitations of the narratives in the news

As we have seen, the five news articles studied here introduce critical lenses to look at the workings and shortcomings of the COMPAS algorithm. However, as sources to learn about this technology, they narrow our understanding by focusing on the technical aspects and ignoring contextual forces. In this section I describe three limitations of the narratives in the news and augment them by considering the power relations where the algorithm is embedded and operates.

First, by focusing on the technology, the news articles do not involve the contexts of creation and use of the algorithm, and thereby avoid placing responsibility and accountability for the racially skewed results. Both Northpointe and the Criminal Justice System are passive actors in these stories, and it is important to bring them to light in order to envision how things can be different.

Secondly, all the news articles center their attention on race, while rendering other possible biases invisible. Since the risk factors that COMPAS uses include socioeconomic status, gender and employment status, as well as a set of personality-related questions, there are other injustices

that are likely to happen. Sonja B Starr is one of the very few scholars who has turned her attention to other forms of discrimination of COMPAS besides race, and argues that determining the future of a person in relation to the factors mentioned above is unconstitutional, since they can become the root of a new set of latent mistakes. For example punishing people because of their conditions of poverty, unemployment or regarding their gender identity (Starr, 2-4).

Third, while the articles by The Atlantic and The Guardian briefly acknowledge that the COMPAS algorithm is one of many risk assessment tools used for years within the criminal justice system, they do not connect this fact with a larger context of unfair incarcerations. Desmarais and Singh analyzed 19 assessment instruments or systems designed for predicting risk of general recidivism and identified 47 instruments designed for specific jurisdictions, declaring that all the instruments "have similar levels of performance" (208), in other words, all of them exhibit the same type of results as COMPAS and therefore the same type of mistakes. This happens because all the tools "are tapping common factors or shared dimensions of risk, even though the instruments utilize different items or have different approaches" (James, 4). By knowing that the COMPAS algorithm has been around for two decades and that it is not the only tool that creates the same effect in society, we can start decentralizing the narratives from this particular technology, in order to see that there is a larger issue of "products that are designed with a lack of careful analysis about their potential impact on a diverse array of people." (Noble, 66). In this case, the news articles do not mention the reality of mass incarceration of people of color in the U.S., which is ultimately the context that this algorithm is aggravating. While mass incarceration is not the focus of this paper, it is indeed necessary to bear in mind that the COMPAS algorithm is part of a larger political project of sending more people to jail, not necessarily as a strategy to manage crime, but rather to control the population (Alexander, 13).

The COMPAS algorithm beyond the technology

By bringing to light actors and context of creation and use, I aim to pose the COMPAS algorithm as a specific enabler of bias, but not the solely responsible for it. There are other enablers in the context of this technology including people, institutions, companies, law enforcement agents and discourses that are perpetuating the already unacceptable inequalities within the criminal justice system. My analysis does not consider the COMPAS algorithm as an object whose essence is completely technological, but rather embraces the contexts that it is part of and recognizes that the technology affects society and culture, in the same way they affect technology as well. As Annany and Crawford have pointed out, Machine learning systems are "sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans" (974).

Analyzing the COMPAS algorithm under these considerations allowed me to envision scenarios for change. On a first level, the moments of use and the roles of the people involved in operating this technology could be modified. For example, by defining what the numerical score produced by the algorithm is used for. There must be a recalibration between what the technology is meant to do, and how it is actually used. The COMPAS algorithm should help judges assign rehabilitation needs, rather than determine length of incarcerations, and the use of this technology should be supervised in order to ensure its consistent use. In addition, after the algorithm runs the risk test, it should complement the numerical score with a report that evinces how the calculation was made along with a mention of the specific aspects of the questionnaire that led to the a particular result, in this way we can take advantage of the complexity of the algorithm by pointing out more precise vulnerabilities, assessing the most relevant rehabilitation needs and allowing for further scrutiny of racial proxies.

It is in the hands of the company Northpointe to revise the initial parameters that the algorithm is built upon, clarifying the instances where the training data contains information that directly correlates to race, poverty, gender, and other criteria that could lead to discrimination. Furthermore, the engineers and programmers should be exposed to education regarding machine biases and the potential harms for society.

This takes us to a third level of influence, the Machine Learning community that is already making some efforts for determining ethical practices for machine learning systems. These efforts should complement the scope of educating the creators of these systems by developing a network of best practices within the private sector and formalizing standards or requirements to deploy these technologies.

The criminal justice system can play different roles for pursuing change. On one hand, they should define clear parameter of fairness, taking a close look at the correlation between previous arrests and risk to recidivate, since the number of arrests in the U.S. is disproportionate for black people, and the impossibility of building a life after prison is in itself a factor that renders people into a higher probability of reoffending (Israni). These nuances should be examined before implementing any technology to predict recidivism. On the other hand, if the intention of the criminal justice system is to define the rehabilitation needs of an offender, not only should the technology be following this principle, but there must be a rehabilitation ecosystem that is ready to take up all the people that should be allocated to them instead of the prison. Furthermore, it is imperative to question the over reliance of the criminal justice system on private companies to manage information, to revise other risk assessment tools that are expressing similar biases like COMPAS, and implement de-bias training for judges and their supervisors.

As a connecting thread amongst these different levels of intervention, there must exist overall communication practices that maintain all the efforts moving towards socially desirable outcomes involving the criminal justice system, the private companies that provide them with products and services, the machine learning community and all the people that operate technologies in a daily basis within this system. These efforts are not needed for a technology to function, but rather to end practices that favor mass incarceration and the perpetuation of social inequalities for specific segments of society.

Conclusion

The narratives about the COMPAS algorithm in the online news articles create a pathway for interpretation that avoids larger structural, historical and sociopolitical forces that have contributed to the creation of this technology and its current use. By emphasizing the analysis in the technology itself, this approach leaves no scope for situating accountability for its racially skewed results. In contrast, approaching the analysis from a contextual perspective reveals how accountability is distributed among the network of actors. More importantly, it highlights the multiple practices that are already biased within the criminal justice system, even without the presence of this technology, to explain that the COMPAS algorithm is not creating a new type of injustice but rather reinforcing long standing inequalities. A deeper look into the history of bias within the criminal justice system, as well as a critical lens to understand race, poverty and punishment is needed for translating this analysis into real changes. This discussion should also be extended to communities of programmers that have the opportunity to envision the long term effects of Machine Learning systems and their underlying systems of classification.

Works cited

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, vol. 20, no. 3, 2016, pp. 973–989., doi:10.1177/1461444816676645.

Angwin, Julia, et al. "Machine Bias." *ProPublica*, ProPublica, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Alexander, Michelle. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2010.

Corbett-Davies, Sam, et al. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." *The Washington Post*, WP Company, 17 Oct. 2016, www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.4be697f8b953.

Desmarais, Sarah L., and Jay P. Singh. *Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States*. Department of Psychology, North Carolina State University and Department of Justice, Psychiatric/Psychological Service, Canton of Zürich, Switzerland, 27 March, 2013

Devlin, Hannah. "Software 'No More Accurate than Untrained Humans' at Judging Reoffending Risk." *The Guardian*, Guardian News and Media, 17 Jan. 2018, www.theguardian.com/us-news/2018/jan/17/software-no-more-accurate-than-untrained-humans-at-judging-reoffending-risk.

Israni, Ellora Thadaney. "When an Algorithm Helps Send You to Prison." *The New York Times*, The New York Times, 26 Oct. 2017, www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

James, Nathan. "Risk and Needs Assessment in the Federal Prison System." *Congressional Research Service*, 10 July 2018.

Mbadiwe, Tafari. *Algorithmic Injustice.* The New Atlantis, no. 54, 2018, pp. 3–28. *JSTOR*, JSTOR, www.jstor.org/stable/90021005.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN: 978-2021200632

Palazzolo, Joe. "Court: Judges Can Consider Predictive Algorithms in Sentencing." *The Wall Street Journal*, Dow Jones & Company, 13 July 2016, blogs.wsj.com/law/2016/07/13/court-judges-can-consider-predictive-algorithms-in-sentencing/

"Practitioner's Guide to COMPAS Core." *Northpointe*, 13 Mar. 2015.

Russell-Brown, Katheryn. "The Academic Swoon Over Implicit Racial Bias." *Du Bois Review: Social Science Research on Race*, vol. 15, no. 01, 2018, pp. 185–193., doi:10.1017/s1742058x18000073.

Starr, Sonja B. "The New Profiling." *Federal Sentencing Reporter*, vol. 27, no. 4, 2015, pp. 229–236., doi:10.1525/fsr.2015.27.4.229.

Yong, Ed. "A Popular Algorithm Is No Better at Predicting Crimes Than Random People." *The Atlantic*, Atlantic Media Company, 29 Jan. 2018, www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.