Artificial Intelligence in Court: An STS Framework Analysis of the vulnerability of Computer-Aided Decision-Making Processes

STS Research Paper

Presented to the Faculty of the

School of Engineering and Applied Science

University of Virginia

By

Xinyue Lin (Mint Lin)

May 4, 2022

On my honor as a University of Virginia student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____Xinyue Lin_____

Approved: _____ Date _____

Introduction

In 2013, 21-year-old black male Bernard Parker was charged with illegal marijuana possession with intent to sell. He was rated a score of 10 out of 10 by Florida's Broward County using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is an algorithm employed by U.S. courts in order to estimate the likelihood of defendant recidivism (Jones, 2018). The outcome of his case was widely discussed, and racial bias was thought to have negatively impacted his COMPAS score due to a comparison of Parker's case with Dylan Fugett's case, a white male who shared similar residential and past criminal charges background with Parker. Specifically, Fugett got a COMPAS score of 3, and therefore as compared with Parker, Fugett should be much less likely to commit subsequent offenses. However, after prison, Fugett was arrested twice in a 2-year window while Parker had none (Jones, 2018). Given their similarities, the difference in their COMPAS scores was likely attributed to race, thereby raising concerns about the unfairness of the COMPAS algorithm (Angwin et al., 2016).

Most literature and technical analysis focus on the algorithmic bias as the primary reason, while few have investigated other factors associated with the entire pipeline of COMPAS usage, starting from the algorithm development to its weight in court. I plan to use the Actor-Network Theory to explain why and how the actors failed to build a stable network through the translational process, helping law enforcers and algorithm engineers understand how each step during the lifecycle of an algorithm influences its fairness altogether. Specifically, I am researching the data collection, algorithm design and evaluation, and algorithm usage processes. I argue that the interactions between and limitations of these actors led to the vulnerability of the COMPAS project. To support my argument, I will use primary sources like the COMPAS

algorithm contract (New York State Division of Criminal Justice Services and NORTHPOINTE Inc. AGREEMENT, 2009) and the Defendant's Questionnaire (Angwin, 2016) to provide evidence. I will also quote scholarly reviews such as Algorithmic Bias in Recidivism Prediction (Khademi & Honavar, 2020) and An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures (Zhang et al., 2014) to provide historical context. I will analyze the A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous *Engineering & Translation* (Cressman, 2009) as a template of the conceptual framework.

Background Information

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a web-based risk and needs assessment instrument that measures criminals' risks of recidivism. It was developed by Northpointe

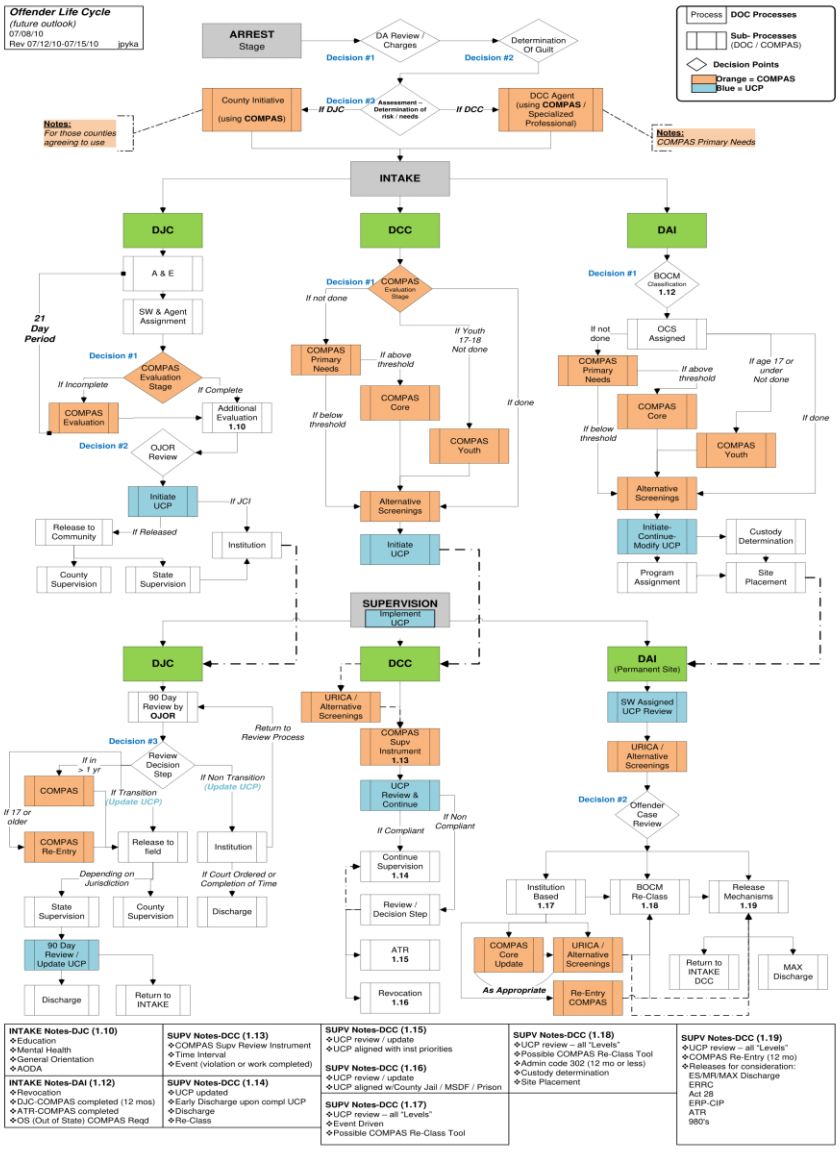

Fig 1. Flowchart of the Offender Life Cycle

(now rebranded itself to “Equivant”) first in 1998 and has been actively updating through time (Equivant, 2019). This algorithm has been used as a decision support system by a few places in the United States, including Wisconsin, Florida’s Broward County, and California (Kirkpatrick, 2017). COMPAS contains two major risk models, the General Recidivism Risk and the Violent Recidivism Risk, whose prediction results serve as reference to judges during sentencing. COMPAS scores range from 1 to 10, with 1 to 4 labeled a low risk of recidivism, 5 to 7 labeled a medium risk, and 8 to 10 labeled a high risk (Equivant, 2019).

The COMPAS score has been relied on heavily in courts. For example, the State of Wisconsin utilizes COMPAS at various decision points during sentencing, as shown in the Figure 1 logical flowchart, where orange blocks imply the use of COMPAS (State of Wisconsin Department of Corrections, 2010).

Literature Review

The performance of the COMPAS algorithm influences the results defendants get, and several scholarly sources have focused on analyzing the accuracy of the COMPAS algorithms by employing causality analysis and comparing COMPAS judgments to human evaluations. For instance, Aria Khademi and Vasant Honavar utilized a scientific tool called FACT, a recently developed measure of algorithmic fairness using causal inference, to analyze whether COMPAS is biased against African American defendants (Khademi & Honavar, 2020). Their results exhibited robust evidence that COMPAS has racial biases against African Americans.

The publication written by Julia Dressel and Hany Farid also assessed the accuracy of COMPAS by comparing COMPAS judgments to people with little law knowledge (Dressel &

Farid, 2018). People from an online crowdsourcing marketplace were randomly drawn as evaluators. According to their discussion, both COMPAS and untrained humans reached a performance ceiling of around 65%, casting doubts on the prediction accurateness of COMPAS.

Additionally, the California Department of Corrections and Rehabilitation has conducted a 3-year large-scale study to investigate the risks of the COMPAS algorithm (Zhang et al., 2014). The findings showed that the predictive power for subsequent criminal offenses is lower than expected. Furthermore, they managed to train a simpler model that takes in only four risk factors, including gender, age, age of the first arrest, and the number of first arrests and can perform no worse than the COMPAS algorithm that makes use of 137 distinct features, implying the inefficiency of the COMPAS model.

While all three scholarly publications discussed above assessed the accuracy of the COMPAS algorithm, which is valuable and insightful, they do not try to explain why and how the accuracy score is obtained as a result of algorithmic design. The next step of the analysis will be to pay attention to factors that play a role in the COMPAS design and the court decision-making process, investigating both the data and the judges. The interaction of human and non-human factors influences the current state of the COMPAS algorithm as well, and I will contribute to the understanding of COMPAS by analyzing these in the following sections.

Conceptual Framework

The Actor-Network Theory (ANT) will be utilized to address why COMPAS is vulnerable as a decision support tool in Bernard Parker's court case. ANT was first developed in the early 1980s by Bruno Latour, Michael Callon, and John Law in an attempt to comprehend the

social aspects behind science (Callon, 1986). ANT assumes that any society is made up of actors and their interactions, where all actors are characterized by their positions in the network as well as their relationships with others. Each network is considered to be dynamically evolving, and any actor -- including "individuals, groups, texts, technical artifacts, organizations, and animals" – can make a difference in the network (Lowe, 2001). In ANT, all actors, including human and non-human ones, are equally powerful and influential in the network (Cressman, 2009). According to Cressman, all actors express agency, and "they can be allies or adversaries of the network – working for or against the interests of the network."

Callon put emphasis on the formation of such a network, developing a concept called "Translation", consisting of four stages: problematization, interessement, enrolment, and mobilization (Callon, 1986). During problematization, the actor defines a problem or a goal and involves other actors to solve or accomplish it. Interessement refers to the stage where network builders lock human and non-human actors to join and collaboratively build the network. Enrolment is the process where strategies are defined and roles are assigned to actors. Mobilization is the final process where actors are ensured to act in favor of collectivities. The network then functions like a black box, where heterogeneous actors act coherently to establish this stable network.

In the analysis that follows, I will begin by identifying the COMPAS network builder. Then, I will name the key actors, both human and non-human involved in the dynamically developing process of the COMPAS. Finally, I will analyze how they contribute to the vulnerability of COMPAS together at different stages and whether their contributions are positive or negative with respect to the common goal of building a just risk evaluation system for defendants.

Analysis of Evidence

During the first stage of the process of translation, Problemitization, the network builders refer to the developer company Northpointe and the New York State Division of Criminal Justice Service. They set a goal to design a fair risk assessment system for the judiciary. Next, during Interessement, the developers recruit actors to join the network. The important actors this paper focuses on include: the developers, the data, the algorithm design, the evaluation metrics, and the judges. Each actor performs its job as scripted and interacts with each other. In this section, I will use primary and secondary sources to first analyze each actor and then draw on their connections and how they shape the development and the current vulnerable state of the COMPAS algorithm.

Northpointe and the New York State Division of Criminal Justice Service

As the recruiters, the New York State Division of Criminal Justice Service (DCJS) and Northpointe Inc were bounded by a contract they signed, whose limitations created difficulties for other actors such as the COMPAS algorithm developers to comply and solve the problem of creating fairness in court. A fair algorithm design requires frequently detecting and fixing coding issues, which is made complicated to achieve due to the lack of code visibility due to signing the contract. Northpointe agreed to develop COMPAS for sentencing usage (New York State Division of Criminal Justice Services and NORTHPOINTE Inc. AGREEMENT, 2009). As the service provider that was profiting based on its product, Northpointe was required to keep its COMPAS algorithm closed-source. Since the code is not visible to the public, experts in relevant technical fields could not critique or give advice on model improvements. Therefore, the only way to find and correct potential bugs was through Northpointe developers, another important actor in the network. The absence of public supervision could result in a higher risk of making

biased algorithmic decisions while putting more burdens on the developers, making them harder to conform voluntarily in the interessement and enrolment stages.

COMPAS algorithm developers

As a result of the contract, developers were recruited by Northpointe Inc as the actor to implement the code that computed the COMPAS score during the interessement process. While the developers attempted to comply with the rules to develop a just judiciary system, they met difficulties introduced by three non-human factors: the data, the algorithm design, and the evaluation metrics, which resulted in a deviation from their scripted actions during the enrolment stage.

Data

The data fed in the COMPAS model largely determined how and what the algorithm will learn to predict. The biased training data could create a partial COMPAS score calculation that was beyond developers' control due to the information extraction technique utilized to collect the data. The input to this algorithm came from the answers to a survey with 137 questions about each defendant's background information (Angwin, 2016). Some questions included the defendant's address, prior arrests, and convictions, whether drugs were available in the defendants' neighborhood, the frequency of the defendant moving residences, and the amount of money the defendant had. The developers had left the defendant's race out of the data, possibly to reduce the chances of the algorithm relying heavily on the race information to make a judgment. However, even though the race information was not explicitly stated, many of the variables COMPAS obtained from the survey as mentioned above had a strong correlation with race.

Consequently, it was difficult to isolate the effects of race on the COMPAS algorithm despite the developers' efforts to correct the mistake, potentially resulting in an algorithmic bias against race.

Additionally, the unbalanced distribution of past criminal records used as the input to the COMPAS algorithm could negatively influence the system's fairness, which the developers could not control. According to Larson and his colleagues, the data input included more crimes committed by black than white defendants (Larson et al., 2016). Among the 7000+ usable criminal records obtained through the Broward County Clerk's Office in Florida, 51.4% of those were marked as the black race, and 34.0% were marked as the white race. The percentage suggested that the data contained considerably more black crimes than white crimes. Florida had one of the strongest open-records laws that allowed the public to obtain crime information (Larson et al., 2016). Even though the data sample did not come from all courts that use the COMPAS algorithm, it was reasonable to suspect that the observable difference in proportions would also be likely observed across different places in the United States of America. Consequently, the unequal distribution of data input, which was not a controllable factor for the developers, could prompt the algorithm to focus too much on black defendants' attributes while ignoring key features for white defendants, potentially resulting in biased results even if the developers had a fair algorithmic skeleton design.

Algorithmic design

By definition of the usage of an algorithmic design, developers would be training and fitting the model on public data and then applying this algorithm to generate a solution for an individual court case, which introduced mistakes by nature (Zhang et al., 2014). According to Sheldon X. Zhang and his fellow researchers, the use of a model trained on public data implied that we were trying to apply results from group analysis on individual cases, which still remained

a challenge. There was a lot of room for improvement regarding reducing the margins of such predictions based on static variables. Indeed, since different individuals could be drastically different, the employment of group-derived analysis would not be able to capture the unique attributes of each defendant and, therefore, will inevitably mispredict to some extent, potentially resulting in unfair judgments.

Evaluation metrics for upgrades

According to the contract signed between DCJS and Northpointe, Northpointe developers were responsible for upgrading the COMPAS algorithm to maintain the software's operation quality (New York State Division of Criminal Justice Services and NORTHPOINTE Inc. AGREEMENT, 2009). Consequently, the evaluation metrics that examined the need for an upgrade could send incorrect signals to developers about when to fix the algorithm system, which impacted the quality system. The error estimation of the COMPAS score was an essential evaluation metric that was obtained by observing the recidivism rate after prison. The data heavily depended on how recidivism was defined and counted. I would describe three factors that influenced how recidivism was calculated below.

Firstly, the definition of recidivism determined the range of crimes we focused on. For instance, the most commonly used FBI description of violent crimes includes "murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault (FBI, 2010)." Consequently, minor offenses that defendants commit would not be counted by definition.

Secondly, the length of time during which the number of recidivism was counted also impacted the crime data that would be collected, impacting the accuracy of the evaluation metrics. Based on the practitioner's guide on COMPAS provided by Northpointe, the correlation

between the COMPAS score and the chances of repeated criminal offenses after the jail was only studied in a two-year window (Equivant, 2019). The 'two years' time restriction was a manually defined time frame. Therefore, if a defendant committed a crime two years after his arrest, he would be recorded as having no recidivism.

Thirdly, a defendant's new offense will be recorded only if the police have pursued his or her crime. Therefore, if a defendant commits a crime not identified or luckily runs away without getting arrested and charged, his or her record of recidivism will still be clean. While escaped or unobserved criminal activities can be hard to prevent, it inevitably impacts Northpointe developers' judgments on whether COMPAS needs an algorithmic fix.

To sum up, the definition of recidivism, the time window for committing crimes, and the unrecorded offenses all impacted the Northpointe developer's judgment of the accuracy of the COMPAS algorithm by manipulating the evaluation metrics for an upgrade, which could delay updates, preventing the developers from developing an algorithm of justice in the interessement and enrolment process.

Court Judges

Court judges were a group of important actors who were served by the COMPAS algorithm and who worked in the network during enrolment by reviewing defendants' information and making sentencing decisions (Larson et al., 2016). However, this type of actor tended to deviate from their scripted actions by being an adversary of the network who worked against the common interests due to the COMPAS score provided for them during sentencing. Any extra information could implicitly influence judges' thinking processes. For instance, if the COMPAS score were an inaccurate summary of the defendant's background information that

exaggerated how dangerous they were, the judge would be more inclined to conclude on a higher offense level, resulting in unfairness. Consequently, the judges could be held partially accountable for an unsuccessful enrolment stage, making it difficult for other actions, such as the developers and the offeror and offeree of the COMPAS contract to accomplish the task of creating justice in court.

Summary

I have shown that the judges were in opposition to the network. However, people might argue that many courts, such as the Wisconsin Supreme Court, require warnings about the algorithmic risk assessments to be presented to judges, meaning the COMPAS score presented during the court should do no harm to the justice during sentencing if not resulted in more positive effects. Indeed, the presentencing investigating reports (PSIs) created for judges must include five written warnings about how the risk scores were calculated, COMPAS's potential inability to identify specific individuals, concerns raised by previous scholarly studies, and the initial purpose of COMPAS (Harv. L. Rev., 2016). These warnings should remind judges to reference the COMPAS scores with cautions. However, this viewpoint is problematic because, according to the theory of Confirmation Bias in psychology, people tend to search for and interpret information to favor their own judgments (Thornhill et al., 2019). While the judges were believed to have received official training on making judgments, if the COMPAS assessment disagreed partially with their sets of beliefs, they might have an implicit tendency to favor the part of the COMPAS results over the rest.

Conclusion

The court case about Bernard Parker raises concerns about the validity of the COMPAS algorithm used to assess defendants of their risks to society. Building upon previous literature that measures the accuracy of the COMPAS model, I proceed further to analyze why and how the COMPAS algorithm is vulnerable as we observe it in its current state. Through employing the Actor-Network Theory, I have identified several crucial actors, namely the contract, the data, the algorithm, the judges, the warning messages of COMPAS scores, the defendants, and the evaluation metrics, who play a role in the lifecycle of the COMPAS algorithm design that is thought to be biased against black defendants. As discussed in the analysis section, all these actors and their interactions directly or indirectly change the effectiveness of the COMPAS algorithm, therefore putting this model in a weak situation. It is worth noticing that some of them are beyond our control. For instance, the contract makes it impossible for the public to examine the algorithm code of this program, defendants who committed uncaught crimes after receiving the COMPAS score will have a recidivism rate of 0, and there is currently no better way to infer about individuals using results obtained from group studies. The potential limitations and room for development can help law enforcers and engineers to get a better understanding of AI mechanisms in law settings, prompting them to design, improve, and act justly in the face of racial biases.

Total Word Count: 3272

References

Angwin, J. (2016). *Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"*. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 25). *What Algorithmic Injustice Looks Like in Real Life*. ProPublica. https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life?token=x1WkMVQkYc7NIixUj9OEexcAEo4anmm4

Brown, Steven D. (2002). "Michel Serres: Science, Translation and the Logic of the Parasite." Theory, Culture & Society 19(3) pp.1-27.

Callon M. (1986). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay, The Sociological Review, 32, 196-223.

Cressman, D. (2009). *A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous Engineering & Translation*. https://summit.sfu.ca/item/13593

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Equivant. (2019, April 4). Practitioner's Guide to Compas Core. Retrieved February 28, 2022, from https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

FBI. (2010). *Violent Crime*. FBI. https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime/violent-crime

Harv. L. Rev. (2016). *State v. Loomis*. https://harvardlawreview.org/2017/03/state-v-loomis/

Jones, R. (2018, August 21). The Siren Song of Objectivity: Risk Assessment Tools and Racial Disparity. *Medium*. https://nacdl.medium.com/from-the-president-the-siren-song-of-objectivity-risk-assessment-tools-and-racial-disparity-fa5ccb0698a5

Khademi, A., & Honavar, V. (2020). Algorithmic Bias in Recidivism Prediction: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(10), 13839–13840. https://doi.org/10.1609/aaai.v34i10.7192

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. https://doi.org/10.1145/3022181

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lowe, A. (2001). After ANT - An illustrative discussion of the implications for qualitative accounting case research. Accounting, Auditing & Accountability Journal, 14(3), 327–351. https://doi.org/10.1108/EUM0000000005519

New York State Dvision of Criminal Justice Services and NORTHPOINTE Inc. AGREEMENT. (2009, Oct 24). https://www.documentcloud.org/documents/20431817-cm00902-northpointe-compas-signed-contract

Thornhill, C., Meeus, Q., Peperkamp, J., & Berendt, B. (2019). A Digital Nudge to Counter Confirmation Bias. Frontiers in Big Data, 2. https://www.frontiersin.org/article/10.3389/fdata.2019.00011.

Washington, A. (2019). How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate. *Undefined*. https://www.semanticscholar.org/paper/How-to-Argue-with-an-Algorithm%3A-Lessons-from-the-Washington/a333464ae4387b3c72a65ededd53aa5eb79c9846

Zhang, S. X., Roberts, R. E. L., & Farabee, D. (2014). An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures. *Crime & Delinquency*, *60*(2), 167–192. https://doi.org/10.1177/0011128711426544

State of Wisconsin Department of Corrections (July 8, 2010), Offender Life Cycle.

https://doc.wi.gov/Documents/AboutDOC/Reentry/offenderlifecycle.pdf