An Actor-Network Theory Analysis of Algorithms in America's Criminal Justice System

STS Research Paper

Presented to the Faculty of the

School of Engineering and Applied Science

University of Virginia

By

Rachel Choi

April 23, 2021

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

Signed: ____

Approved: _____________________________________ Date ________________________

Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

Artificial intelligence is a wide-ranging branch of computer science in which the focus is on building smart machines that are capable of performing tasks that typically require human intelligence. As advancements in machine learning and deep learning are creating a paradigm shift in almost every sector of the tech industry, artificial intelligence is creating a huge influence on the way we live, work, travel, and do business in the 21st century.

The influence of AI technology can also be seen in the criminal justice industry. Artificial Intelligence is widely used throughout the criminal justice system in the United States. The most commonly used are pretrial risk assessment algorithms, also called as risk assessment tools, which are designed to predict a defendant's future risk for reoffending. They influence judgments about guilt or innocence, bail, and sentencing. However, the algorithms are largely hidden from public view, and many critics have raised concerns that the results may be a source of bias (Angwin et al., 2016).

Throughout this paper, I will use Actor-Network Theory (ANT) to show the fragility of risk assessment tool in the US justice system. I will explore the risk assessment algorithm and show how it is biased to raise awareness of possible unfair sentencing. Drawing on ANT, I will focus on the risk assessment tool and argue that there are potential problems of algorithmic risk assessments used in the judicial system.

Background

Risk assessment instruments (RAIs) are designed to predict a defendant’s future risk for misconduct. These predictions inform high-stakes judicial decisions, such as whether to incarcerate an individual before their trial. For example, an RAI called the Public Safety

Assessment (PSA) considers an individual’s age and history of misconduct, along with other factors, to produce three different risk scores: the risk that they will be convicted for any new crime, the risk that they will be convicted for a new violent crime, and the risk that they will fail to appear in court. A decision-making algorithm translates these risk scores into release-condition recommendations, with higher risk scores corresponding to stricter release conditions. RAIs influence a wide variety of judicial decisions, including sentencing decisions and probation and parole requirements (Chohlas-Wood, 2020).

Risk assessment tools are used in almost every state in the U.S. They are usually used in pre-trial, although they exist at sentencing, in prison management, and for parole determinations. There are also specific risk assessment tools for different functions in the criminal justice system, such as domestic violence risk or juvenile justice risks, with the understanding that different factors are used in those contexts than in a general criminal risk or violent criminal risk of rearrest or re-offense (Chohlas-Wood, 2020). As criminal justice algorithms have come into greater use at the federal and state levels, they have also come under greater scrutiny. Many criminal justice experts have denounced risk assessment tools as opaque, unreliable, and unconstitutional.

Literature Review

Many scholars have analyzed risk assessment tools that are used in the justice system to measure the advantages and disadvantages of risk assessment tools. According to some scholars, RAIs have the potential to bring consistency, accuracy, and transparency to judicial decisions as it can bring consistent decisions compared to human influenced decisions. For example, Jung et al. did a simulation on the use of a simple checklist-style RAI that only considered the age of the defendant and their number of prior failures to appear. The scholars found that judges in an undisclosed jurisdiction had widely varying release rates (from roughly 50% to almost 90% of individuals released). The scholars noted that if judges had used the proposed checklist-style assessment model to determine pretrial release, decisions would have been more consistent across cases, and they would have detained 30% fewer defendants overall without a corresponding rise in pretrial misconduct.

In addition, other studies have found additional evidence that statistical models can outperform unaided human decisions as the decision are not solely influenced by the judges but are based on statistical models. Structured risk assessment tools are believed to reduce the likelihood the evaluator's estimate of an offender's reoffending risk will be influenced by bias, which is a systematic error in reasoning or logic that occurs as the result of the automaticity with which the human mind processes information based on expectations and experience (Tversky and Kahneman, 1974). A good example of this phenomenon is confirmation bias, which occurs when attention is drawn to evidence that supports a favored scenario or outcome, while evidence that weakens or contradicts the preferred hypothesis is discounted or ignored altogether. In fact, research suggest that criminal investigators and police trainees tend to view evidence such as witness statements, DNA evidence, and photo evidence as less credible and reliable if the evidence contradicts their beliefs about the guilt of the suspect (Ask and Granhag, 2007). The benefits of statistical assessment have compelled many jurisdictions across the country to implement RAIs.

However, in parallel with the expansion of RAIs across the country, many scholars have shown that there is a potential bias behind the results of the risk assessment tools. One analysis was done on the Ohio Risk Assessment System (ORAS), which was developed as a statewide system to assess the risk and needs of Ohio offenders in order to improve consistency and facilitate communication across criminal justice agencies. The goal was to develop assessment tools that were predictive of recidivism at multiple points in the criminal justice system. Specifically, assessment instruments were to be developed at the following stages: pretrial, community supervision, institutional intake, and community reentry (Latessa et all., 2010). According to the study, it shows some gender bias by resulting in lower risk levels for females. Figure 1 presents the percentages of offenders that recidivated for each risk level of the RT by gender. The results indicate that both male and female groups experienced increasing rates of recidivism for each risk level. For males. 21 percent of low-risk cases were rearrested, 50 percent of moderate-risk cases were rearrested, and 64 percent of high-risk cases were rearrested. The r value of .29 indicates that the RT does a good job of distinguishing between low, moderate, and high-risk cases. For females, however, only six percent of low-risk females were arrested, while 44 percent of moderate-risk cases were arrested, and 56 percent of high-risk cases were arrested. The large r value of .44 is likely a result of the substantial difference between low- and moderate-risk females. However, the result suggests that findings for females should be taken with caution as it can include gender bias on the results of the RAIs.

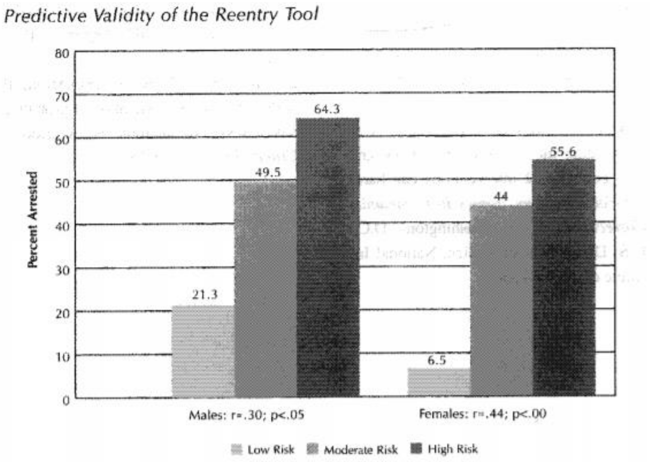

Figure 1. Predictive Validity of the Reentry Tool

Another scholarly research that has been done in regards to the risk assessment tool is an analysis of a tool made by Northpointe, Inc., COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Northpointe created risk scales for general and violent recidivism and for pretrial misconduct. According to the COMPAS Practitioner's Guide, the scales were designed using behavioral and psychological constructs of very high relevance to recidivism and criminal careers. COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and other jurisdictions (Larson et al., 2016). The research group conducted an analysis on COMPAS and found 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. The study was done on more than 10,000 criminal defendants in Broward County, Florida. The research group compared the recidivism risk categories predicted by the COMPAS tool to the actual recidivism rates of defendants in the two years after they were scored, and found that the score correctly predicted an offender’s recidivism 61 percent of the time, but was only correct in its predictions of violent recidivism 20 percent of the time. In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants and 63 percent for black defendants) but made mistakes in very different ways. It misclassified the white and black defendants differently when examined over a two-year follow-up period. It was found that black defendants were often predicted to be at a higher risk of recidivism than they actually were. Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent). Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism; White violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.

The first two sources show how the risk assessment tools can help by making decisions that are not human influenced, which can help make impartial judgments. In contrary, the source authored by Latessa states that there is gender bias related in the risk assessment tools, and the second source authored by Larson states that there is racial bias in the risk assessment tools. I will deploy ANT to show the possibility of bias in the risk assessment tools.

Conceptual Framework

The risk assessment network can be analyzed using the Actor-Network Theory (ANT). ANT examines complex relationships between human and nonhuman actors that come together to create a dynamic network. The main idea is that scientific knowledge is an effect of established relations between objects, animals, and humans engaged in scientific practices (Detel, 2002).

Within the framework of ANT, Callon introduces the concept of translation. Translation describes the process of forming and maintaining an actor-network. The stages of translation consist of problematization, interessement, enrolment, mobilization, and black-box. In problematization, network builders identify a problem and the actors needed to solve the problem. The essential actors must be shaped to serve the network's goal, which becomes the groundwork to establish the network. In interessement, the network builders seek out additional actors, both human and non-human, to join the network and contribute to the primary actors' goals. In enrollment, network builders assign roles to the recruited actors, and the recruited actors accept and perform their assignments. In mobilization, network builders maintain their role of representing the network and speak for the network's initial goal. In black-box, all the actors function as a coherent stable network.

According to Callon, a network can fail at any step of the way, especially when an actor fails to act in good faith for the network's original goal. These actors are rogue actors. If a rogue actor does not play the role scripted for it by the network builders, the entire network becomes vulnerable and unstable. The idea of rogue actor is important when analyzing the risk assessment network, as the risk assessment tools acted as a rogue actor by producing a biased result, which fails to contribute to the network's primary goal of producing fair rulings.

Analysis

Network Formation

To analyze the fragility of the risk assessment network, it is important to provide the background. The network involves multiple human and nonhuman actors that are interconnected. The network builders include technologists and legal experts who have created the risk assessment tools. The translation of ANT is formed with the network builders. In the stage of problematization, the technologists and legal experts identify the problem of slow sentencing process due to the massive number of defendants. For the network builders, the goal would be to design an algorithm that would effectively analyze the recidivism of defendants to help expedite the sentencing process. Potential motivation for these network builders would be the desire to create a fair ruling system for defendants and faster decision making for the judges, and money for the company they work for. In the stage of interessement, technologists would seek out additional actors, such as private companies and judicial officers, to work towards their common goal. In the stage of enrollment, network builders would assign roles to the actors, such as asking private companies and judicial officers for sponsorship to create the algorithm. In the stages of mobilization and black-box, the actors in the risk assessment network will maintain their roles and work towards their common goal of creating an accurate risk assessment tool.

One of the nonhuman actors in the network is the Wisconsin v Loomis (2016) case, in which the Wisconsin Supreme Court held that a trial court's use of an algorithmic risk assessment in sentencing did not violate the defendant's due process rights even though the methodology used to produce the assessment was disclosed neither to the court nor to the defendant (Ossei-Owusu, 2017).

Another nonhuman actor in the network is COMPAS, a risk assessment tool made by Northpointe, Inc. This tool was used in the ruling of Wisconsin v Loomis (2016), which influenced the sentencing. In preparation for sentencing, a Wisconsin Department of Corrections officer produced a pre-sentence investigation Report (PSI) that included a COMPAS risk assessment. COMPAS assessments estimate the risk of recidivism based on both an interview with the offender and information from the offender's criminal history. The methodology behind COMPAS is kept as a secret, and only the estimates of recidivism risk are reported to the court. At Loomis's sentencing hearing, the trial court referred to the COMPAS assessment in its sentencing determination and, based in part on this assessment, sentenced Loomis to six years of imprisonment and five years of extended supervision (Harvard Law Review, 2017).

One of the human actors in the network is the defendant, Eric Loomis. In February 2013, Eric Loomis was found driving a car that had been used in a shooting. He was arrested, and pleaded guilty to eluding an officer and no contest to operating a vehicle without the owner's consent. When Loomis was sentenced for eluding the police, the judge told him he presented a "high risk" to the community and handed down a six-year prison term. The judge said he had arrived at his sentencing decision in part because of Mr. Loomis's rating on the COMPAS assessment. Loomis challenged the judge's reliance on the COMPAS score, appealing on the criteria used by the COMPAS algorithm, which is proprietary and result is protected, and on the differences in its application for men and women.

Other additional actors include panel of judges who support the automated system as they help judges rely on more than educated guesses in deciding what happens to the defendants. In fact, in a recent poll by the National Judicial College of 369 judges, a clear majority (65%) agreed that artificial intelligence can be a useful tool for combatting bias in bail and sentencing decisions, but it should never completely replace a judge's discretion (American Bar Association, 2020).

Other actors also include some technologists and legal experts who are skeptic about the risk assessment algorithms. The technologists raise concerns on how machine-learning algorithms use statistics to find patterns in data. Thus, if historical crime data is fed to the algorithm, it will pick out the patterns associated with crime. However, these patterns are statistical correlations, not causations. Thus, the risk assessment algorithms will turn correlative insights into causal scoring mechanisms (Hao, 2019). Meanwhile, legal experts argue that the algorithms will make the legal system more incomprehensible and data-based, as courts have relied more on the automated tools when making decisions in the last few years (Re & Solow-Niederman, 2019). Actors also include community activists in American Civil Liberties Union (ACLU) and Black Lives Matter (BLM). The activists in ACLU argue that human prejudices can be baked into these tools because the machine-learning models are trained on biased police data

(Larson & Schmidt, 2014). Similarly, activists in BLM argue that the risk assessment scores are measured with known sources of bias, such as “race” and “gang affiliation” (Sentencing Project, 2015). They argue that these features produce results that are biased towards their race.

Conclusion

Throughout the paper, I have used ANT framework to outline the dynamics of the risk assessment network. I have shown the fragility of the actor network due to unreliability in the risk assessment algorithms, and I have shown how this directly impacts the justice system. In the case of Wisconsin v Loomis (2016), Justice Ann Walsh Bradley rejected Loomis’s due process arguments. Justice Bradley found that the use of gender as a factor in the risk assessment served the nondiscriminatory purpose of promoting accuracy and that Loomis had not provided sufficient evidence that the sentencing court had actually considered gender. Moreover, as COMPAS uses only publicly available data and data provided by the defendant, the court concluded that Loomis could have denied or explained any information that went into making the report and therefore could have verified the accuracy of the information used in sentencing. However, Justice Bradley added that judges must proceed with caution when using such risk assessments. To ensure that judges weigh risk assessments appropriately, the court prescribed both how these assessments must be presented to trial courts and the extent to which judges may use them. The court explained that risk scores may not be used “to determine whether an offender is incarcerated” or “to determine the severity of the sentence.” Therefore, judges using risk assessments must explain the factors other than the assessment that support the sentence imposed (Harvard Law Review, 2017).

The case of Wisconsin v Loomis (2016) shed light on the fragility of risk assessment tools, and warns the use of algorithm in the justice system. It urges caution in testing the results and eliminating any prejudices in the algorithms. As we are rushing into the world of tomorrow with big-data risk assessment, it is important for us to properly vetting, studying and ensuring that we minimize a lot of these potential biases in the data. This is especially important for Enhancing the Fairness and Effectiveness of the Criminal Justice System. The case also highlights a broader national discussion about how law enforcement officials use predictive data, including deciding which streets to patrol, identifying people at risk of being shot and calculating the likelihood of recidivism.

Word count: 3104

References

American Bar Association. (2020, February 16). The good, bad and ugly of new risk-assessment tech in criminal justice. Retrieved October 20, 2020, from https://www.americanbar.org/news/abanews/aba-news-archives/2020/02/the-good--bad-and-ugly-of-new-risk-assessment-tech-in-criminal-j/

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias. Retrieved October 7, 2020, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ask, K., & Granhag, P. A. (2007). Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology, 37(3)*, 561-591.

Callon M. (1986) The Sociology of an Actor-Network: The Case of the Electric Vehicle. In: Callon M., Law J., Rip A. (eds) Mapping the Dynamics of Science and Technology. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-07408-2_2

Chohlas-Wood, Alex. "Understanding Risk Assessment Instruments in Criminal Justice." *Brookings*, Brookings, 17 June 2020, www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice.

Detel, W. (2002, November 02). Social constructivism. Retrieved March 30, 2021, from https://www.sciencedirect.com/science/article/pii/B008043076701086X

Hao, K. (2019, January 21). AI is sending people to jail-and getting it wrong. Retrieved September 17, 2020, from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Harvard Law Review. (2017, March 10). Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. Retrieved April 03, 2021, from https://harvardlawreview.org/2017/03/state-v-loomis/

Heaven, W. (2020). Predictive policing algorithms are racist. They need to be dismantled. Retrieved September 17, 2020, from https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

Jung, J., Concannon, C., Shroff, R., Goel, S. and Goldstein, D.G. (2020). Simple rules to guide expert classifications. Journal of the Royal Statistical Society: Series A (Statistics in Society). doi:10.1111/rssa.12576

Larson, E., & Schmidt, P. (Eds.). (2014). *The Law and Society Reader II*. NYU Press. Retrieved October 8, 2020, from JSTOR.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. Pro Publica. Retrieved October 20, 2020, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Latessa, E. J., Lemke, R., Makarios, M., & Smith, P. (2010). The creation and validation of the ohio risk assessment system (oras). *Federal Probation*, *74(1),* 16-22.

Ossei-Owusu, Shaun. "Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing." *Harvard Law Review*, 10 Mar. 2017, harvardlawreview.org/2017/03/state-v-loomis/.

Re, R. & Solow-Niederman, A. (2019). Developing Artificially Intelligent Justice. Retrieved September 17, 2020, from https://law.stanford.edu/wp-content/uploads/2019/08/Re-Solow-Niederman_20190808.pdf

The Sentencing Project. (2015). Eliminating Racial Inequality in The Criminal Justice System. Retrieved October 8, 2020, from JSTOR.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: *Heuristics* and biases. *Science, 185(4157),* 1124-1131.

Wagner, P. & Sawyer, W. (2020, March 24). Mass Incarceration: The Whole Pie 2020. Retrieved October 31, 2020, from https://www.prisonpolicy.org/reports/pie2020.html