Implementing Machine Learning in Criminal Profiling, Reviewed Within the Context of the Association of Computing Machinery (ACM) Code of Ethics and Professional Conduct

Abstract

Machine learning (ML) provides many opportunities to improve efficiency and productivity over human-only efforts. The ability to process large amounts of diverse data points in near real time offers potential benefits that might otherwise go untapped.

Notably, one principal truth remains since the advent of the first computer, "garbage in, garbage out". In the case of ML this relates to the quantity and quality of "training" allocated to the implementation of ML. Training of ML applications, conducted by humans, introduces both bias and error.

This literature review examines scholarly papers, articles and case studies to determine if the state of ML/AI as implemented for criminal suspect profiling aligns with the ACM Code of Ethics. Review targets two popular use cases:

- Predicting terrorism activities/selecting subjects for watch lists.
- Generating predictive scores used by many jurisdictions in determining bail bond amounts.

Keywords:

Machine Language, Criminal Profiling, Terrorism, Bias, Ethics, Bail, Recidivism.

Introduction

The revelations by NSA whistleblower Edward Snowden in 2013 brought global attention to the potential "harm" that can be done by well-intentioned programs such as PRISM when used beyond an intended constraint. Most reasonable people would agree that the need to feel and be safe from terrorist activity carries considerable weight in determining how much privacy should be relinquished to ensure security. In the United States, these types of needs are subjected to scrutiny by the FISA court, a secretive process that provides a venue for judicial review but cloaks the process from all but the bare minimum required for processing the case. However, as revealed by Snowden and other cases since, the process is not infallible. At the foundation of all these ethical quandaries is one of the top enabling technologies, ML.

ML has also reached prolific use within the criminal profiling industry. Whereas the NSA program PRISM spied on the whole world collectively, and then used meta data to discover targets, criminal profiling applications are most frequently targeted against specific individuals. In many cases the individuals are not informed of the data collection nor their resulting score. Yet judges rely on these scores to determine eligibility or conditions for bail. This literature review examines whether the creators of these applications used may have violated the ethical code of the ACM.

Types of Literature Selected

Literature for review was researched using established sources of scholarly and peer reviewed papers and articles. These included IEEE, researchgate.org, springer.com, *et al*. Only a few find space in this

review, but many contributed to the author's general knowledge of the subject. No single piece of literature was excluded or discounted based on a divergence in conclusions, but rather all tended to support the consensus. Where the author applies critical review or alternative conclusion, citations in support of the criticality are presented.

Use Case 1. Machine Learning to Combat Terrorism

The primary literature under review for this use case is Machine Learning Against Terrorism: How Big Data Collection and Analysis Influences the Privacy-Security Dilemma. Verhelst *et al*, (2020) points out that the "diverse characteristics of terrorist attacks" make the application of ML ineffective. The ability to capture ever-larger pools of data do not improve the algorithms as the diversity grows at a similar rate to the data pool. This can be attributed to the broad diversity of terrorist motivation, methodology, and target selection. Terrorism comprises differing motives, methods, and recruiting tactics, influenced by nations or regions, religion, ethnicity, and radical politics. The overwhelming amount of big data dilutes and confounds the ML training rather than improving it. (Verhelst *et al*, 2020)

Critical examination of Privacy Principles - Privacy is defined and valued differently based on diverse nationalities, ethnicities, and personal life experiences

As the program manager and principal investigator of a large research project funded by the U.S. Government, I authored and oversaw the implementation of policies and procedures for protection of participant's personally identifiable information. During one visit by dignitaries from an Asian Country, I explained some of those policies and was met by blank stares from the delegation. The concept of personal privacy outside of the home was not part of their cultural norm, and they couldn't grasp what the fuss was all about. But still, nearly everyone has some expectation of privacy. And when that line is crossed people can feel violated.

Critical examination of the premise that this use is ethical because - Machine learning applications for combatting terrorism tend to target groups of people rather than individuals, at least at the highest level of data collection

Having one's information collected as part of a group may seem less concerning for some as it "isn't personal" …. that is until one ends up on a watch list because of a loose connection to some fringe group to whom they don't ascribe. Now it's personal, but hard to overcome. Devereaux (2019) inn reporting on Elhady v Kable (2019), describes how the U.S. District Court for the Eastern District of Virginia found the Terrorist Screening Database was unconstitutional, as it denied listees their constitutional right to freedom of movement without "due process". Reasoning cited in the decision included that information used in the selection of persons for addition to the list, comprise "algorithms" that could or are biased by human "training" data that are almost certainly biased.

Critical examination of the premise that no implicit rights are denied based on this use of ML - Errors or bias in ML applications for combating terrorism may impact constitutionally protected activities

Elhady v Kable (2019) also acknowledged that processes used in screening for the list included monitoring protected activities like freedom of speech, assembly, movement, and freedom to seek redress. The Court further held that inclusion on the list denied plaintiffs of their liberty interests in international travel, interstate travel and being free from Government stigmatization as a terrorist.

Critical examination of ML reliability for this use - The irregularity of terrorism data makes it unreliable for ML training

ML relies on collected data which the algorithm has been trained to interpret. Verhelst, et al, *(*2020) cite three known problems in ML theory:

- Class Imbalance: Foley (2020) describes the issue as resulting from "trying to predict something that doesn't happen very often", in which case, your data on the occurrences is greatly outweighed by the data on when it doesn't happen.
- The Curse of Dimensionality: Yiu (2019) explains this boils down to when your data has too many features. In predicting terrorism with ML, you have an extremely large number of diverse data points: nationality, ethnicity, religion, age, social status, economic conditions, peer influence, travel history, etc. Conversely, the limited number of actual terrorism acts leave the quantity of data vs. data produces unreliable conclusions.
- Spurious correlations: Are 2 or more events or variables that seem related but aren't. In predictive analysis such correlations if used to train ML produce skewed outcomes. For example, a small sampling of suicide car bombings may include data that three out of four involved Toyota cars. As Toyotas are prolific in most of the world, this correlation is not useful in predicting the type of vehicles to be expected in future car bombings. It is reasonable that a greater sampling of data would indicate a wider distribution of vehicle types.

Use Case 2. Machine Learning to Predict Criminal Risk and Recidivism

The primary literature under review for this use case is Machine Bias. There's software used across the country to predict future criminal activity. And it's biased against Blacks (Angwin, et al, 2016). This study, while written for a news source, is conducted by a team of highly regarded, full-time investigative reporters. The publisher, Propublica, was formed in 2010 for the specific purpose of providing in-depth, investigative reporting. They only use employee staff, and their reporting is vetted by reputable news organizations around the world. The researchers raise issues of both machine bias and misuse of the resulting data for criminal prediction scores.

The application of ML in this use case is especially concerning

Whereas use case 1 impacts were mostly limited to privacy, convenience, and reputation, the impact of bias in use case 2 affects victims in a much deeper way. The scores assigned via ML criminal Risk Assessment Tools (RATs) can continue to follow them for years and across multiple facets of life. These include pretrial release vs. confinement, bail amount and conditions, sentencing, incarceration conditions, and parole eligibility. In the U.S., 49 States employ some type of RAT for pretrial risk assignment and sentencing advice. (Pretrialrisk.com, N.D.)

Critical examination of the user's claim that errors occurred at equal rates across racial groups - The margin of error evident in this use case is especially concerning but the impacts by race class are reprehensible.

Angwin, et al, (2020) found that only 20% of those predicted to commit violent crimes within the following 2 years did so (the benchmark on which the RAT tool in question is based). 80% of the people who received higher bail amounts, or were refused bail and confined until trial, and many of which received longer sentences because of their RAT scores, were assigned those scores based on flawed and

biased ML applications. Furthermore, while the RAT vendor accurately claims that the application erred equally across racial groups, the impacts of the errors are not equal. For Whites, the errors were highly skewed toward false negatives, meaning they received lower risk scores but often reoffended at a higher rate than predicted. In contrast, errors for Blacks constituted a much higher rate of false positives, meaning they consistently scored as higher risk but were less likely to reoffend within the 2 years following.

Angwin, et al, (2020) concluded that:

> *"In forecasting who would re-offend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.*
>
> - *The formula was particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at almost twice the rate as White defendants.*
> - *White defendants were mislabeled as low risk more often than Black defendants."*

Critical examination of the lack of validation for this use - There is little substantive research available which addresses the validity of the algorithms used in criminal predictive profiling with ML

Desmarais (2013) examined 19 different risk methodologies used in the United States and found that "in most cases, validity had only been examined in one or two studies" and that "frequently, those investigations were completed by the same people who developed the instrument". And while they intended to review studies in the U.S., examining the existence of racial bias in these applications, their only statement was limited to "The Data do not exist".

While more recently articles and scholarly papers have been written on the topic, no significant experimental design research has been documented. This may be attributable to the fact that the RATs in existence are proprietary, and the producer/owners have refused to share their algorithms for study.

Critical examination of the available body of research on this use - Weaknesses in literature reviewed

Author experienced some disappointment in the lack of peripheral viewpoints that may have provided additional insight and credence to conclusions. For example, use case 2 literature presented some limited examples of research subjects that were of non-black race but obviously of another minority group. Data on that group's treatment within the study parameters would have been valuable.

Author's critical review of this use of ML against the ACM - Is ML's use for this purpose aligned with ACM Code?

Several Principles of the ACM Code of Ethics and Professional Conduct (ACM.org, 2018) may be in play here, but the following are most relevant:

1.2 Avoid harm

> Harm means negative consequences and those have been well established throughout. The code also assigns responsibility for reporting and attempting to resolve harm when it occurs. Verhelst, *et al* (2020) cited judges who were interviewed and responded that they would not have used the COMPAS score as the basis of bail or sentencing decisions

had they known of the inherent bias in the algorithm. This implies that the producers / distributors of the product knew and withheld that information.

1.4 Be fair and take action not to discriminate.

Many of the ML products utilized have a come under scrutiny in recent years and yet producers continue to hide behind intellectual property protections to prohibit an evaluation of their product's compliance with the ACM code and suitability for use in the purpose under which it is marketed. (Kehl & Kessler, 2017; Hannah-Moffat, 2019; and Barabas *et al*., 2017)

3.7 Recognize and take special care of systems that become integrated into the infrastructure of society.

These ML applications were specifically designed for use within the justice system and 3.7 implicitly applies. There is a duty to monitor for appropriate use and misuse.

Conclusions

Author concludes:

- This research is important as the impact of poorly designed or misused ML applications for analysis and prediction of future criminal activity can be severe and affect both individuals and whole classes of subjects, including minority or disadvantaged groups.
- Additional research is needed in this area as the existing literature is limited in both quantity and breadth of scope. Specifically, experimental design research should be conducted to address the validity and reliability of the algorithms used in ML for predictive analysis of future criminal activity. "Definitive data do not exist" (Desmarais, 2013).
- The development of ML applications for use within the justice system and/or predictive modelling of future criminal behavior is subject to the ACM Code. Education and training should be a topic of discussion by the ACM and institutions of higher learning (ACM.org, 2018),

References:

ACM.org (2018) ACM Code of Ethics and Professional Conduct. Available from: https://www.acm.org/code-of-ethics [Accessed February 2022]

Angwin, J, Larson, J, Mattu, S, Kirchner, L, in ProPublica. May 23, 2016. Available from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed February 2022]

Barabas, C, Dinakar, K, Ito, J, Virza, M, Zittrain, J (2017) Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. Cornell University. arXiv:1712.08238 Available from: https://arxiv.org/abs/1712.08238 [Accessed March 2022]

Desmarais, S, Singh, J, and Van Dorn, R. (2013) Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review - Singh - 2013 - Behavioral Sciences &; the Law - Wiley Online Library Available from: https://onlinelibrary.wiley.com/doi/10.1002/bsl.2053 [Accessed March 2022]

Devereaux, R., (2019) SECRET TERRORISM WATCHLIST FOUND UNCONSTITUTIONAL IN HISTORIC DECISION. *In The Dispatch* Available from https://theintercept.com/2019/09/06/terrorism-watchlist-lawsuit-ruling/ [Accessed March 2022]

Elhady v Kable Available from https://int.nyt.com/data/documenthelper/1689-terror-watchlist-ruling/75cd50557652ad0bfa2a/optimized/full.pdf#page=1 [Accessed March 2022]

Foley, 2020 Strategies for Imbalanced Data Daniel Foley | Towards Data Available from: https://towardsdatascience.com/machine-learning-and-class-imbalances-eacb296e776f [Accessed March 2022]

(Hannah-Moffat, 2019), Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates - Kelly Hannah-Moffat, 2019 (sagepub.com) Available from: https://journals.sagepub.com/doi/10.1177/1362480618763582 [Accessed March 2022]

(Kehl and Kessler, 2017), Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041 [Accessed March 2022]

Pretrialrisk.com (N.D.) Outputs: Risk Scores - Mapping Pretrial Injustice. Available from: https://pretrialrisk.com/the-basics/pretrial-risk-assessment-instruments-prai/outputs-risk-scores/ [Accessed March 2022]

Verhelst, H.M., Stannat, A.W. & Mecacci, G. (2020) *Machine Learning Against Terrorism: How Big Data Collection and Analysis Influences the Privacy-Security Dilemma*. *Sci Eng Ethics* **26,** 2975–2984. https://doi.org/10.1007/s11948-020-00254-w [Accessed February 2022)

Yiu, 2019 The Curse of Dimensionality. Why High Dimensional Data Can Be So… | by Tony Yiu | Towards Data Science. Available from: https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e [Accessed March 2022]

Bibliography: (reviewed and helped form the author's general consensus but not specifically quoted or summarized)

Dethlefsen, J, (2019) The Ethics of Machine Learning and Discrimination in Viterbi Conversations in Ethics (usc.edu). Available from: https://vce.usc.edu/volume-3-issue-2/the-ethics-of-machine-learning-and-discrimination/ [Accessed March 2022]

Rigano, C, 2019, "Using Artificial Intelligence to Address Criminal Justice Needs," NIJ Journal 280, January 2019, https:// www.nij.gov/journals/280/Pages/using-artificialintelligence-to-address-criminal-justice-needs.aspx [Accessed February 2022)

Rudin, C, Wang, T, Wagner, D, and Sevieri, R, (2013), Predictive Policing: Using Machine Learning to Detect Patterns of Crime. Available from: https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/ [Accessed February 2022]

Winters, B, (N.D.) AI in the Criminal Justice System. Available from: https://epic.org/issues/ai/ai-in-the-criminal-justice-system/ [Accessed March 2022]

Zavrsnik, A, (2019), Algorithmic justice: Algorithms and big data in criminal justice settings in the European Journal of Criminology. Available from: https://journals.sagepub.com/doi/10.1177/1477370819876762ce [Accessed February 2022)