Examining Behavioral Detection Systems on a Varying Range of Social Integration

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Benny Bigler-Wang

Spring 2024

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

Advisors

William F. Stafford, Jr., Department of Engineering and Society

Introduction

Artificial Intelligence (AI) has revolutionized various domains, including the development of Behavioral Detection Systems. These systems are designed to interpret and respond to human behaviors in diverse contexts. The underlying goal of training a machine learning model is to derive the ground truth meaning of a problem which can then be used to determine solutions for future, novel instances. The problem with developing a model to distinguish certain actions or social behaviors is that we ourselves do not know what the ground truth is, or we maybe know that there is no specific ground truth at all for a certain scenario. For these reasons, developing a Behavioral Detection System for a problem that we do not know the specifics for or is not specific enough itself will be increasingly difficult especially when wanting to minimize bias. Currently, there is research in producing models that minimize bias and discrimination using methods that aim for fairness using certain heuristics and mathematical abstractions (Dwork, C. et al., 2012; Zemel R. et al., 2013), but it is hard to ensure these qualities especially while maintaining a high level of accuracy for the model. While the current technology to ensure fairness is still mostly under research, a big part in reducing discrimination in Behavioral Detection Systems requires better diversification of training data and more robust data augmentation (Ferrara E.). This means even if we are able to develop methods that are able to precisely evaluate the fairness of certain models we also need to change the way we collect data and select or augment it to ensure that we are capturing a diverse, representative dataset. A part of this involves diversifying the teams that are tasked with AI development and evaluation to bring more contrasting perspectives that will better identify and correct biases in the process (Holstein, K., 2019).

Behavioral Detection Systems

The term "Behavioral Detection Systems" is a loose term that has been talked about more in recent literature and encompasses a broad spectrum of meanings across different fields. For the purpose of this discussion, I define it, broadly, as a system employing machine learning techniques to identify specific behaviors. In cybersecurity, for instance, this definition extends to include both behavioral and anomaly detection, where the former targets specific behaviors as evidence of an attack, while the latter detects deviations from a baseline behavioral pattern established during training. I will be diving deeper into a few different scenarios of Behavioral

Detection System use in the following sections, exploring the degree to which privacy concerns and potential discrimination may arise in relation to "social integration" as a conceptual tool to analyze various applications of Behavioral Detection Systems. By examining these issues, I aim to show a correlation between how socially integrated a Behavioral Detection System is to the potential for discrimination within the problem's application, formalizing the need for more research into fairness in AI and the continuous evaluation of how we use AI to solve our world's problems. In doing so, I delineate a spectrum of social integration that refers to the degree to which the application of a Behavioral Detection System interacts with and impacts human society.

Literature Review

The literature on the integration of artificial intelligence (AI) across various domains, including cybersecurity, healthcare, and government surveillance, reveals a growing interest in leveraging AI technologies for improved outcomes and efficiency. Kumar et al. (2023) highlight the increasing use of AI in the cybersecurity landscape, particularly in threat hunting, vulnerability management, and network security. Their work emphasizes three primary challenges associated with AI integration in cybersecurity: human adversaries, AI-cyber attacks, and ethical considerations. While providing a broad overview of AI applications in cybersecurity, Kumar et al. also discuss the need for future growth, integration, and regulation of AI, underscoring the importance of addressing ethical implications, as echoed in other scholarly works.

In the healthcare sector, Ahuja AS (2019) explores the integration of AI, specifically Behavioral Detection Systems, to monitor patients' behaviors and detect subtle changes indicative of various health conditions. The potential of these systems to revolutionize healthcare by offering universal access and personalized treatment plans regardless of socio-economic backgrounds is highlighted. This is possible due to the capability of these AI systems to optimize and expedite the healthcare diagnostic and treatment process at the expense of relatively low cost. However, Murdoch (2021) raises concerns regarding data privacy, particularly in the context of healthcare AI solutions. Murdoch emphasizes the significant issue of data privacy, especially regarding personal medical data, which is considered highly private and legally protected, suggesting the need for differential privacy methods in healthcare and possibly

extending it to cybersecurity systems safeguarding healthcare data. The lack of regulation and oversight in this area also raises concerns about potential discrimination in diagnostic systems that have access to patient healthcare data which can lead to suboptimal or even detrimental outcomes for the patients..

Saura, J. R. (2022) delves into the government's use of AI in behavioral analysis of citizens, discussing concepts such as surveillance capitalism and behavioral data science, which are the widespread collection of personal data by corporations and the study of issues regarding human behavior, respectively. Saura conducts a systematic literature review on how governments utilize AI and its implications for privacy. The ease with which governments can access citizen data, including economic, social, and health data, to predict behavior raises significant privacy concerns. Collective behavior analysis and predictions pose a threat to user privacy, suggesting potential limitations on the use of Behavioral Detection Systems in certain contexts due to privacy risks.

Overall, the literature review highlights the multifaceted implications of integrating Behavioral Detection Systems across different domains. While these systems hold promise for improving outcomes and efficiency, ethical considerations, data privacy concerns, and potential discriminatory practices underscore the need for comprehensive regulation and oversight in the development and implementation of Behavioral Detection Systems.

Discussion of Cases

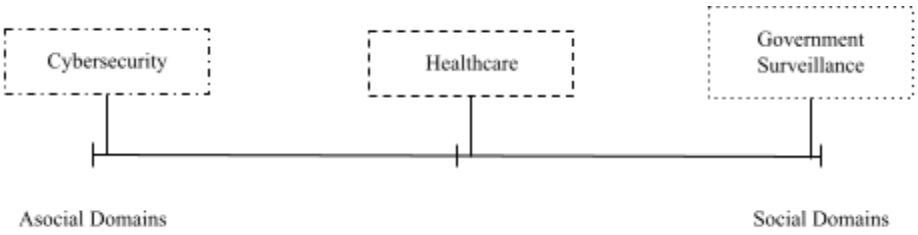

Figure 1. Social Integration Spectrum

We can look at multiple cases of Behavioral Detection Systems being used in different scenarios with each having varying levels of integration of its application into society. The following cases are arranged to highlight both extreme ends of the social integration spectrum shown in Figure 1 and then one case that can be considered the average between the other two

which can be seen as taking a “goldilocks” structure. Each of these cases represent different fields of work and therefore have a lot of differences contextually. I labeled the cases in the cybersecurity field to be on the lower end of the socially integrated spectrum, the cases relating to healthcare applications to be in the middle of the socially integrated spectrum, and the government surveillance case to be on the higher end of the socially integrated spectrum. As we will see, the more connected the application of the problem is to society the more considerations and types of biases are introduced and compounded on top of one another.

In the realm of cybersecurity, Behavioral Detection Systems represent a pivotal shift in ensuring the security of digital assets without having major social implications. Unlike the traditional forms of malware detection that rely on signature-based or heuristic-based methods, behavioral-based detection focuses on identifying unusual patterns or deviations in user behavior that may indicate the presence of malware. These systems are becoming increasingly more useful at detecting novel cyber-attacks or new evolving malware strains, commonly referred to as zero-day threats, which are threats that were previously unknown to the public. Older systems using signature-based detection are ineffective against zero-day threats as attackers can easily evade detection by modifying the code or using obfuscation techniques (O’Kane, P., 2011). Therefore adopting a behavioral approach combats one of the major weaknesses of signature-based detection systems.

Working at Vali Cyber, I got to evaluate their existing behavioral detection engine through offensive penetration testing. They are developing a Linux security agent, ZeroLock, that utilizes a behavioral detection engine. Diving deeper into a specific attack vector, cryptojacking, that the engine defends against, we are able to see how advantageous behavioral detection models are at protecting machines. Cryptojacking is the unauthorized use of other people’s devices to mine for cryptocurrency, essentially stealing other people’s processing power. Vali Cyber created a feature set of the distribution of bitwise operators used in the assembly code of certain cryptojacking malware strains and then trained an SVM model to recognize the pattern behind these specific attempts at cryptojacking. After training the model, they were able to achieve a 99.98% accuracy rate at detecting cryptojacking on future testing data. This application of machine learning for behavioral detection was extremely effective because cryptojacking produces a very unique distribution of bitwise operators, specifically a lot of ANDs in the assembly code. Therefore in this context, the Behavioral Detection System that was designed to

detect cryptojacking behavior on a machine was integrated very minimally into society as cryptojacking is a problem that exists in the domain of the computer systems and computer networks.

The integration of artificial intelligence into healthcare appears very alluring as there is great potential to improve patient care and treatment outcomes. Behavioral detection systems, utilizing AI algorithms, can monitor patients' behaviors and detect subtle changes that may indicate the progression of various health conditions. These systems have the potential to revolutionize healthcare by providing universal access and personalized treatment plans tailored to an individual's needs regardless of their socio-economic background (Ahuja AS, 2019). However, amidst the excitement surrounding AI-driven healthcare solutions arise concerns regarding the potential for the misuse of patient data, biased outcomes, and discriminatory practices.

It is now fairly easy for technology companies to collect daily activity data from smartphones and smartwatches or through other sources and obtain insights that may have been considered private otherwise. It is said that these practices are out of line and do not guarantee HIPAA compliance (Na, L., 2018). Therefore even though these systems can provide an incredible amount of utility for healthcare providers, it opens up the doors for numerous concerns regarding data privacy. Additional privacy concerns are also present due to the progress made in model inversion attacks that attempt to reverse engineer models in order to extract information about specific individuals in the dataset. This can pose a significant threat to models that are trained on sensitive data, as it is in this instance in the healthcare field. It was proven by Zhang Y. (2020) that a model's predictive power and vulnerability to inversion attacks are highly correlated and that canonical forms of differential privacy had very little effect on ensuring privacy. More robust frameworks and differential privacy methods must be developed before we implement Behavioral Detection Systems on healthcare data.

Recent studies have also highlighted the presence of racial bias in algorithms widely used in the healthcare industry which make it challenging to deliver the previously mentioned equitable delivery of healthcare services to all individuals. In one such algorithm, it was seen that black patients were given the same risk score as white patients even though they were sicker,

resulting in black patients being recommended for extra care half as much as they should have (Obermeyer, Z., 2019). This bias occurred due to the historical disparity of black people and white people in healthcare spending represented in the data collected, with black people usually spending less. Therefore, biased algorithms may exacerbate existing healthcare disparities by perpetuating stereotypes and reinforcing systemic inequalities. Health care is a field of work that requires social interaction between the patient and healthcare provider. Thus, the problem of diagnosis is one that is mainly scientific but reliant on the healthcare provider's judgment and the system in place to treat the patient. Similar to what was said by Murdoch, B. (2021), the regulation of AI in healthcare has fallen behind the existing technologies causing a lot of oversight in the field. Addressing the issue of bias in Behavioral Detection Systems requires a multifaceted approach that involves critical examination of the data sources, algorithmic design, and decision-making processes underlying these systems.

The utilization of Behavioral Detection Systems in government surveillance introduces complex socio-technical dynamics that intersect with issues of privacy, civil liberties, and the potential for discrimination. The goal of these systems is to analyze individual's behaviors to detect potential threats to national security or public safety. This analysis can be done on data gathered from various sources such as surveillance cameras, social media, platforms, and communications to create detailed profiles of individuals' actions.

This indiscriminate collection and analysis of peoples' data raise concerns about the invasion of privacy and the potential for mass surveillance. As has already been seen in many authoritarian states such as the People's Republic of China, the Soviet Union, and North Korea, large-scale surveillance systems have been used to monitor public spaces online or in the real-world. This is not exclusive to authoritarian states as seen with the Snowden incident in 2013 and what followed, with many of the world's liberal democracies moving towards government surveillance, incorporating Behavioral Detection Systems that can automatically capture and alert specific patterns of behavior in society. A review of literature on government surveillance concluded that while there were 11 main uses of AI to enhance the interactions between citizens, organizations, and public services, the review also uncovered 8 key topics of concern regarding citizens' privacy (Saura, J. R., 2022), such as behavioral predictions, data privacy law and regulation, risk of behavior modification, etc. Biases inherent in the data

collected, such as racial, gender, or socio-economic biases, can also lead to compounding discriminatory practices in government surveillance. This bias is as much of a result of biased surveillance practices as it is a result of big data collection and modeling. Tackling the problems of surveillance detection systems poses a great challenge due to the scale and complexity of the task. In a sense, government surveillance is at the extreme end of social integration because the algorithm is being trained on an individual's interactions within society hence surveying society itself. The tradeoffs between security and the ethical implications of implementing such an idea would need to be formalized in great detail by politics and ethics experts, and even then would need to be heavily scrutinized and continuously updated to effectively enforce current legislation and morals of society. Additionally, transparency and accountability in data collection and analysis would need to be ensured in order to mitigate the risk of discrimination in government surveillance practices.

Analysis

The discussion of the aforementioned cases provide a general overview of the introduction of Behavioral Detection Systems into various fields, highlighting the spectrum of social integration, from cybersecurity to healthcare to government surveillance.

In the realm of cybersecurity, Behavioral Detection Systems represents a significant advancement in threat detection, particularly in combating zero-day threats that traditional signature-based systems struggle to identify. The case study of Vali Cyber illustrates the effectiveness of Behavioral Detection Systems in detecting specific cyber threats, such as cryptojacking, through machine learning algorithms. This application demonstrates minimal social integration as the focus remains primarily on protecting digital assets within computer systems and networks, with limited direct societal implications beyond safeguarding data and infrastructure. Implementations of Behavioral Detection Systems that would fall into this same category of minimal social integration should have minimal restrictions or regulations placed upon them, for the faster AI can be used to solve these problems the better off we will be.

On the other hand, the integration of Behavioral Detection Systems into healthcare initiates a myriad of ethical and privacy considerations. While there is great potential for improving patient care and treatment outcomes through AI-driven healthcare solutions, concerns

regarding data privacy, biased outcomes, and discriminatory practices emerge. The case studies highlight the risks associated with the collection and analysis of sensitive patient data, as well as the presence of historical racial bias in algorithms used for healthcare decision-making. This underscores the need for a multifaceted approach to address bias in Behavioral Detection Systems, involving critical examination of data sources, algorithmic design, and decision-making processes to ensure equitable healthcare delivery. In domains such as healthcare, where the problems are fairly socially integrated and yet brought about through some form of scientific/logical reasoning, Behavioral Detection Systems can provide a lot of utility indiscriminately with correct execution. These types of problems are complex, but with the proper use of expert knowledge in that particular domain and well-defined ethical regulation used for such systems, application of such systems could be developed effectively in the near future.

The discussion extends to the application of Behavioral Detection Systems in government surveillance, where the stakes are significantly higher in terms of privacy invasion and civil liberties. Government surveillance systems aim to analyze individuals' behaviors to detect potential threats to national security or public safety, often through indiscriminate data collection from various sources. This raises concerns about mass surveillance and the potential for discriminatory practices fueled by biases inherent in the collected data. The case study highlights the need for transparency, accountability, and formalized trade-offs between security and ethical implications in government surveillance practices to mitigate the risk of discrimination and uphold civil liberties. When it comes to formalizing the ethics and regulations surrounding Behavioral Detection Systems, including government detection cases, a multidisciplinary approach is necessary. While government surveillance raises significant ethical concerns, it serves as a critical case study to examine the social implications of Behavioral Detection System applications. Government agencies, alongside ethicists, legal experts, technologists, and community representatives, should collaborate to develop comprehensive guidelines that address the ethical, legal, and societal implications of Behavioral Detection System implementation.

In her book, Discriminating Data, Chun suggests that “we need […] to understand how machine learning and other algorithms have been embedded with human prejudice and discrimination, not simply at the level of data, but also at the levels of procedure, prediction, and

logic" (2021). Truly understanding how machine learning works will most likely take a long time, so before we reach that point of understanding we need to be vigilant in how we implement Behavioral Detection Systems in order to reduce bias. One approach to mitigating bias involves limiting the application of AI to socially integrated problems where the consequences of bias are less pronounced. This implies that until our fairness and differential privacy models reach a certain level of sophistication, it may be prudent to restrict the use of AI to domains where the social implications are more manageable. By focusing on problems that are less socially integrated, we can minimize the potential harm caused by biased AI algorithms while simultaneously advancing our understanding of fairness and privacy in AI research. Defining the level of social integration poses another challenge in the context of AI ethics and regulation. Social integration can be understood as the degree to which AI systems interact with and impact human society. However, quantifying this concept requires careful consideration of various factors, including the scope of the AI application, its potential consequences on individuals and communities, and the broader societal context. Establishing a metric or proxy to measure social integration can aid in evaluating the ethical implications of AI systems and informing regulatory frameworks.

Citations:

1. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).
2. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In *International conference on machine learning* (pp. 325-333). PMLR.
3. Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*. 2024; 6(1):3. https://doi.org/10.3390/sci6010003
4. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-16).
5. Kumar, S., Gupta, U., Singh, A. K., & Singh, A. K. (2023). Artificial Intelligence. *Journal of Computers, Mechanical and Management*, *2*(3), 31–42. https://doi.org/10.57159/gadl.jcmm.2.3.23064
6. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019 Oct 4;7:e7702. doi: 10.7717/peerj.7702. PMID: 31592346; PMCID: PMC6779111.
7. Murdoch, B. (2021). Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Medical Ethics*, *22*(1). https://doi.org/10.1186/s12910-021-00687-3
8. Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2022). Assessing behavioral data science privacy issues in Government Artificial Intelligence Deployment. *Government Information Quarterly*, *39*(4), 101679. https://doi.org/10.1016/j.giq.2022.101679
9. O'Kane, P., Sezer, S., & McLaughlin, K. (2011). Obfuscation: The Hidden Malware. *IEEE Security & Privacy*, *9*(5), 41–47. https://doi.org/10.1109/msp.2011.98
10. Na, L., Yang, C., Lo, C. C., Zhao, F., Fukuoka, Y., & Aswani, A. (2018). Feasibility of reidentifying individuals in large national physical activity data sets from which protected

health information has been removed with use of machine learning. *JAMA network open*, *1*(8), e186040-e186040.

11. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D. (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 253-261).
12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.
13. Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. MIT press.